Geoinformatica (2016) 20:879–921
DOI 10.1007/s10707-016-0247-0

The big data of violent events: algorithms for association analysis using spatio-temporal storytelling

Raimundo F. Dos Santos Jr.[1] · Arnold Boedihardjo[1] · Sumit Shah[2] · Feng Chen[3] · Chang-Tien Lu[2] · Naren Ramakrishnan[2]

Received: 7 December 2014 / Revised: 15 November 2015 /
Accepted: 2 February 2016 / Published online: 27 February 2016

Abstract This paper proposes three methods of association analysis that address two challenges of Big Data: capturing relatedness among real-world events in high data volumes, and modeling similar events that are described disparately under high data variability. The proposed methods take as input a set of geotemporally-encoded text streams about violent events called "storylines". These storylines are associated for two purposes: to investigate if an event could occur again, and to measure influence, i.e., how one event could help explain the occurrence of another. The first proposed method, *Distance-based Bayesian Inference*, uses spatial distance to relate similar events that are described differently, addressing the challenge of high variability. The second and third methods, *Spatial Association Index* and *Spatio-logical Inference*, measure the influence of storylines in different locations, dealing with the high-volume challenge. Extensive experiments on social unrest in *Mexico* and wars in the *Middle East* showed that these methods can achieve precision and recall as high as

✉ Raimundo F. Dos Santos Jr.
raimundo.f.dossantos@usace.army.mil

Arnold Boedihardjo
arnold.p.boedihardjo@usace.army.mil

Sumit Shah
sshah@vt.edu

Feng Chen
fchen5@albany.edu

Chang-Tien Lu
ctlu@vt.edu

Naren Ramakrishnan
naren@vt.edu

[1] U.S. Army Corps of Engineers, Washington, DC, USA

[2] Virginia Tech - Computer Science Department, 7054 Haycock Rd, Falls Church, VA 22043, USA

[3] State University of New York (SUNY) at Albany, Albany, NY, USA

80 % in retrieval tasks that use both keywords and geospatial information as search criteria. In addition, the experiments demonstrated high effectiveness in uncovering real-world storylines for exploratory analysis.

Keywords Spatial-temporal systems · Entity relationship modeling · Social media networks · Spatial and physical reasoning · Semantic networks · Big Data

1 Introduction

Violent events are often the byproducts of complex factors of various natures (financial, political, and religious). For a violent event to take place, the right mix of signals must come together in order to elicit reaction. Take as an example Fig. 1, which depicts some of the locations of the *Poll Tax Riots* of Great Britain in 1990. Social unrest broke out after the government enacted a flat-rate tax on each adult. But before those acts of violence occurred, other developments led up to them: activists organized protests at *Trafalgar Square*, police closed a few of London's *Underground* stations, transit was rerouted in some streets, and shops closed in certain areas. The key idea here is that violent events tend to be associated with other spatially and temporally-related nearby processes, which are prevalent in Big Data. These processes are composed of any number of constituent parts that, when identified properly, can help uncover the final event.

While the above example is not surprising (after all, protests can frequently lead to riots), acts of violence are not always transparent. The *Montreal Stanley Cup Riot* of 1993, for instance, developed quickly as the crowd celebrated a win, and had no apparent reason to engage in violence, when in fact it did. Violent events can take on many characteristics, four of which are observed in the above example and stated below:

1. **event cascading:** single developments provide little insight into the overall event. On their own, *street closures* are not alarming. But when combined with other developments, such as *gathering of protesters* and *closed shops*, a much bleaker picture begins to emerge;
2. **event propagation:** developments evolve in spatial regions through nearby areas, fading into the distance. Shops, for instance, close doors near the event, but not far away from it;

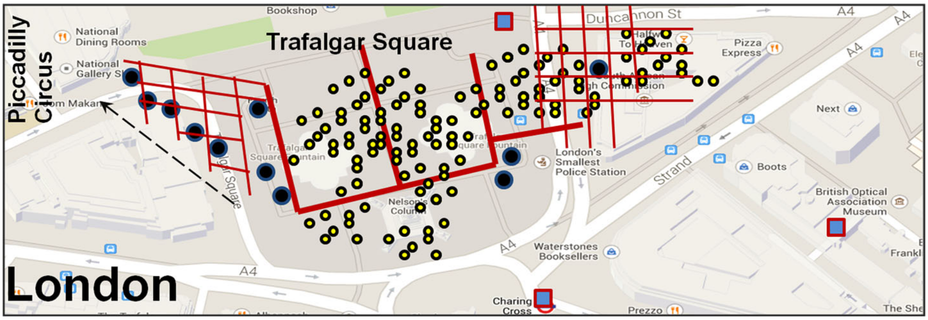


Fig. 1 Approximate spread of the *Poll Tax Riots* of London in 1990. Red lines represent street closures around *Trafalgar Square*. Yellow dots denote concentration of protesters. Squares are closed subway stations, and black dots show locations of reported riots propagating north towards *Piccadilly Circus*

3. **event sequencing:** the temporal sequence in which developments occur is essential to explain facts. *Disruption in transportation*, for example, commonly takes place after protesters have gathered, but less frequently before;
4. **event interaction:** developments represent interactions among entities: police try to contain protesters, rioters throw stones, looters attack shops, etc. Some interactions provoke strong reactions, while others do not.

This study targets two Big Data challenges: determining which events influence one another in high data volumes; and identifying instances of related events described in disparate formats which could reoccur in time. One interesting question is whether such an event (highly observed, but described in various ways), can be associated with others based on previous knowledge of seemingly related developments. In other words, would it be possible to infer looting at London's *Piccadilly Circus* from protests that took place earlier at *Trafalgar Square*? Making such determinations has proven elusive even with the most advanced reasoning systems available today [33].

While making associations between events has been an art as much as a science, the connection strength between two events can be estimated by expanding the four characteristics mentioned above to the following hypothesis: an event can be identified by the **constituent parts** that lead to it, observing their **spatial propagation** and **temporal ordering**, and taking into account their **semantic interactions**. The goal of this study is to reason over spatio-temporal sequences of developments in Big Data (i.e., storylines) that can lead to other events, and provide a numerical view of their associations. For a focused discussion, violent events are used as a case study.

Figure 2 gives an example of what this work entails. The figure shows two event sequences **A** and **B** across a short timeline (31 March, 1990). **Sequence A** is composed of three developments taking place along the day: *Whitcomb St.* is blocked, protests occur at *Trafalgar Square*, and *Charing Cross* station is closed. They culminate in looting in the vicinity of *Piccadilly Circus* at 7 pm. **Sequence B** has three different events in different locations, but also leads to the same looting at *Piccadilly Circus*. Given the two sequences (and possibly others), the goal is to give each sequence a numerical quantification of how they are associated with the looting. One would like to say that **Sequence A** is associated

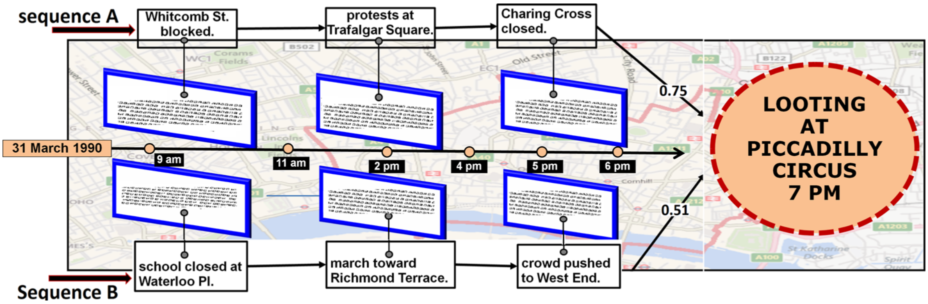

sequence A
Whitcomb St. blocked.
protests at Trafalgar Square.
Charing Cross closed.
0.75
LOOTING AT PICCADILLY CIRCUS 7 PM
31 March 1990
9 am
11 am
2 pm
4 pm
5 pm
6 pm
0.51
school closed at Waterloo Pl.
march toward Richmond Terrace.
crowd pushed to West End.
Sequence B

Fig. 2 Example of association from spatio-temporal storytelling on two event sequences. Sequence A explains the looting at *Piccadilly Circus* as an implication of the blocking of *Whitcomb St.*, protests at *Trafalgar Square*, and closing of *Charing Cross*. Sequence B, alternatively, relates the same looting with a school closing at *Waterloo Pl*, a march to *Richmond Terrace*, and the pushing of the crowd to *West End*. The 0.75 and 0.51 values indicate the beliefs with which sequences A and B are respectively associated with the looting. For its higher value, sequence A has a higher level of association to the looting than sequence B

with the looting at *Piccadilly Circus* with a certain value (0.75), while **Sequence B** is associated with the same looting with a lower value (0.51) than **Sequence A**. Thus, **Sequence A** is more tightly associated with the looting than **B**. These values can be either a probability or an index, two approaches explored later. As noted earlier, the above sequences represent streams of information that *tell a story*, and thus we frame this problem as one of *storytelling*, which is explained below.

Broadly speaking, *storytelling* is the process of connecting entities through their characteristics, actions, and events [29] in order to create meaningful streams of information. In the *Poll Tax Riots* example above, a possible storyline would be the sequence $\boxed{\textbf{\textit{activists}}} \xrightarrow{organize} \boxed{\textbf{\textit{protest}}} \xrightarrow{containedby} \boxed{\textbf{\textit{police}}} \xrightarrow{closed} \boxed{\textbf{\textit{streets}}}$, where entities {activists, protest, police, streets} are connected through semantic relationships {organize, containedby, closed}, and tagged with a location and timestamp. *Information retrieval* and web research have studied this problem, i.e., modeling storylines from documents and search results, and linking documents into stories [7, 9, 14] (the terms *stories* and *storylines* are used interchangeably). A violent event can be viewed as a vector of three important dimensions: the **spatial regions** where entities interact; **temporal coherence** which dictates the proper ordering of developments; and the **interactions** that lead to social outcomes. This study enforces these three dimensions and focuses on spatio-temporal *storytelling* related to violent events, presenting the following contributions:

1. **Designing spatio-temporal methods to analyze events**: Violent events become more humanly-understandable in short spans of space and time. For this purpose, this paper proposes two methods: a *Spatial Association Index* (*SAI*), which measures the relatedness among nearby events and addresses the high-volume challenge of Big Data; and a distance-based variation of *Bayesian Inference* (*DbB*), which relaxes the notion of similarity between storylines and deals with the high-variability challenge of Big Data.
2. **Reasoning with Spatio-logical Inference**: Key to understanding violent events is to differentiate their relevant components from unimportant ones. This work proposes *Spatio-logical Inference* (*SLI*) to find the likelihood that an outcome will occur and deals with the high-volume challenge of Big Data. In this manner, analysts can focus on fewer happenings rather than thousands (or millions) of uninformative developments.
3. **Analyzing events in disparate data formats**: The high variability of Big Data imposes no constraint on how information may be received. On that account, this research utilizes both structured and unstructured data sources, pre-processes them as storylines, and provides association reasoning that can be helpful for practical use regardless of their vocabularies.
4. **Performing extensive experiments over disparate datasets**: Because Big Data comprises various formats, this work performs extensive experiments with different data sources. Namely, it uses *Twitter* data and the *Global Database of Events, Language, and Tone* (*GDELT*), the latter being a well-established dataset of events from where we target conflicts in the Middle East and other parts of Asia.

As noted above, this paper presents three methods of associating analysis. The first, *Distance-based Bayesian Inference*, calculates the probability that an event will occur again. In contrast to traditional *Bayesian Inference*, it allows events of different natures (e.g., “riot” and “demonstration”) to be considered the same for frequency calculation purposes, which is more realistic in the real world. The second, *Spatial Association Index*, measures the influence of one event in one location to another event in another location. Its goal is to find if the former has any correlation to the latter. The third and last method, *Spatio-logical Inference*, investigates if two or more events have a strong (or weak) association with another event such that the first ones can explain the last one. Please see Sections 3.3, 4.1, and 4.2 for an explanation of each method along with numerical examples.

In summary, the first method finds reoccurrences of events, the second measures their influence on one another, and the third investigates if some events can explain others. Clearly, they perform different tasks, and thus are not competitors. However, they can be complementary to one another. Since the first method finds frequent events, these events could be used as input to the second method, which calculates levels of influence in different spatial regions. Further, the most influential events could be used as input to the third method as a way to understand how they propagate in space and time. It is also important to notice that some of these methods may be appropriate for some tasks, but not relevant to others. For example, in applications in which every event is just as important as the next, all events may have to be investigated. In these cases, finding the most frequent ones by Bayesian Inference may not be needed. Finding which ones are more influential (second method), however, may be important to differentiate the ones that cause more impact in the environment. With that in mind, the above methods are designed such that they operate independently from one another. The remainder of this paper is organized as follows. Section 2 describes related work and points their differences to the proposed approaches. Section 3 explains the framework and definitions used in this paper. The discussion on association analysis continues in Section 4, detailing the proposed approaches, while extensive experiments are presented in Section 5. A conclusion is finally given in Section 6.

2 Related work

Storytelling comprises a set of analytical tasks that can be performed in many ways. It can be best described as a platform of knowledge exploration for fact finding, association discovery, and decision-making. The work proposed in this paper, therefore, spans many areas of expertise, from graph analysis to geographic networks. This research best lines up with the approaches described below.

2.1 Storytelling and connecting the dots

The phrase ‘storytelling’ was introduced by Kumar et al. [14] as a generalization of *redescription mining*. At a high level, *redescription mining* takes as input a set of objects and subsets defined over those objects with the goal of identifying objects described in two or more different ways. In [9], Hossain et. al. develop this idea to connect two unrelated *PubMed* documents where connectivity is defined based on a graph structure, using the notions of hammocks (similarity) and cliques (neighborhoods). This work was generalized to entity networks in [8] and specifically targeted for use in intelligence analysis. This class of work represents *traditional storytelling* approaches that do not take into consideration the geospatial features present in the data.

In the realm of frequent pattern mining, research related to this paper comes from *Cascading Spatio-Temporal Pattern Discovery (CSTP)*, proposed by Mohan et al. [20]. *CSTP* identifies partially-ordered subsets of event types that are colocated and sequential. The goal of this approach is not to perform storytelling per se, but its focus on event association is a significant step in that direction. With modifications, CSTP can be a valuable tool complementary to this paper with respect to the proposed *Spatial Association Index* and to the rule generation of the *Spatio-Logical Inference*.

Connecting the dots-type approaches focus on document linkage rather than entity connectivity. They apply textual reasoning as a strong facet of the targeted methods. Link strength utilizes the notion of *coherence* across documents, which is proposed by [25]. Stories are modeled as chains of articles, where the appearance of shared words across documents help establish their relatedness. Extending that work, the authors also propose related methods to generate document summaries, i.e. *Metro Maps*, in [27] and [26], which target scientific literature. Overall, *connecting the dots* methods rely on the abundance of robust content, which cannot be assumed with *Twitter* and *GDELT* data. Thus, *connecting the dots* is less than ideal for such data feeds.

Regarding the approaches discussed above, this paper strives to incorporate some of their strengths to be applicable in a Big Data scenario. It uses *Ripley's K function* as a density metric to shrink the data space from millions of entities to a range in the thousands. It also uses probabilities of the most frequent events such that not all of them need to be investigated. And working under the general assumption that Big Data does not impose a limit on document numbers, spatio-temporal storytelling does not operate at the document level. Rather, it constrains the views at the entity level regardless of the documents in which they are described.

This study focuses on violent events. However, there are several domains of inquiry that can be related to storytelling's event sequences in different scenarios. In *Transportation Planning*, for instance, location recommendation, whether for safety reasons or tourism purposes, has been a popular area of investigation. The work of Zhang et al. [32] proposes an approach (*LORE*) to exploit sequential influence on location recommendations that incrementally mines patterns from location sequences and predicts the probability of a user visiting a certain location. This type of itinerary planning is also targeted by Bolzoni et al. [1] by adding category information to points of interest (POI). Storytelling provides some of the same features by connecting locations to interesting real-world developments. They differ, however, in that the former are concerned with identifying interesting locations, whereas storytelling focuses on finding interesting events across different places. A more subtle use of storytelling can be done in the field of communications. Emergency responders, as an example, can enhance their use of daily information posted on social media. Liu et al. [17] devised an application to detect traffic events based on terms posted on *Twitter*. These terms can be described as the interactions between two entities in a road accident, for example. In [2], Bouros et al. exploit the concept of influence to identify users that can impact a large number of other important users within a given spatial region, which could be useful in understanding viral marketing and other developments. The above areas are only some of the many fields which can benefit from the type of association analysis that this research contributes, and serves to demonstrate storytelling's wide levels of applicability.

2.2 Inferencing and forecasting

While the goal of this study is not to perform forecasting, it entails association analysis, which has a certain affinity with forecasting approaches. Some authors prefer the terms 'event prediction' while others speak of 'causality'. One such work proposed by Radinsky et al. reasons over the causes of events described in news articles [23]. They present an algorithm that takes as input a causality pair to find a causality predictor based on entities.

They depart from this paper's approach, which does not compare entity attributes, but rather investigates association in close spatial proximity.

Another work worth mentioning is prediction from textual data described in [21, 22]. The authors propose to capture the effects of an event by propagating it through a hierarchical model, namely an *abstraction tree*, that contains events and rules. In this paper, a rule-based method is proposed, but does not rely on a trained model that stores rules for subsequent use. It compares events, which may be viewed as nodes with weights and does not depend on the extensive availability of entity attributes.

The above discussion hints at the importance of *Bayesian Inference* in forecasting. Among classical methods, it is one of the strongest foundations for *cause-effect* relationships. *Bayesian Inference* in its traditional form, however, is challenging for a few reasons: (1) it needs many instances of the same events to occur in like sequences to establish certainty; (2) without modification, it does not consider subjective criteria, such as behavioral knowledge. Things "are" or "are not"; (3) it does not take into account spatial reasoning. Every element, no matter where they reside, are regarded equally. In terms of violent events, these three aspects represent challenges that must be dealt with. For this reason, this study does utilize *Bayesian Inference* as an association method, but does not solely rely on it. This study will also propose three other approaches that can handle some of the Big Data challenges in specific domains of application.

2.3 Link analysis

Often relying on graphs as a modeling abstraction, Link Analysis observes the evolution of entities in space and time [5, 19] and the identification of patterns [4, 6]. Link analysis has become popular because Big Data can be better viewed as a graph, rather than as a collection of disconnected documents. The goal of link analysis, however, is not to explore stories or do association mining. Rather, it is an attempt to quantify changes in entities and manage relationships, which leads to the notion of ranking.

Ranking has been popularly applied to web pages since the seminal works of Brin and Page [3] and Kleinberg [12]. The former computes the importance of a web page based on its links and an initial damping factor. The latter also considers the page's links, but is dependent on an initial query that generates a *root set*, and is augmented by other pages that point to the *root set*.

Within the same family of the above approaches, there have been other proposed methods, such as the *Indegree Algorithm* [18] and the *HITS Algorithm* [12]. The former considers the *popularity* factor as a ranking measure while the latter introduced the notion of *hub and authority*. In terms of *storytelling*, both of these types of ranking would be challenging since popularity is too subjective a concept, and there is no clear-cut way to determine which entities would be authorities and which would be hubs. In general, link analysis takes a graph as input and operates on it. Our work, however, goes back one step, and generates the graph from Big Data as a prior requirement. It then performs the necessary analysis on the graph.

Differences Each of the above research fields provides benefits to the various tasks involved in *storytelling*. The proposed work in this paper, for instance, requires geolocation of entities as it relies on a spatio-temporal model. This brings significant benefits to handling Big Data, since space and time can serve as limiting filters in high data volumes, and spatio-temporal coherence can help identify similar events described disparately. Link analysis and Connecting the Dots can help identify important relationships. Given the benefits of each method, this paper does not show competing approaches. Rather, it presents complementary techniques that demonstrate how storylines can be a valuable analysis tool, addresses some of the volume and variability concerns of Big Data, and covers a spatio-temporal niche which remains largely untapped.

3 On relating violent events

This section provides preliminary information that describes the proposed research. Section 3.1 discusses justifications for the ideas in this study. Section 3.2 introduces the analysis framework and definitions used throughout the remainder of the paper. The first proposed method begins in Section 3.3.

3.1 Reasoning over violent events

The main objective of this study is to investigate how storylines are related to one another using different association strategies. In terms of Big Data, this study is concerned with effectiveness of the results, more so than their efficiency. In other words, storylines must be generated and associated using the most amount of information in a timely manner. However, finding true storylines that reflect real-world developments takes priority over finding every possible storyline, many of which may have no truth to them.

A key consideration here is determining if a sequence of initial events has any relationship to a subsequent final event, in which case the former would be deemed associated with the latter. Alternatively, it can also be thought of in terms of propagation, i.e., whether one storyline influences another. Note that this study refrains from claiming that events are being forecast, predicted, or detected. Nor does it claim that one event causes another, as those assertions can be strongly misleading and highly uncertain. Without heavy data analysis and strong supporting evidence, this type of **causality** is extremely difficult to demonstrate [13].

Consider, for example, the *Boston Marathon Bombings* of 2013, which was reported in millions of data feeds, and which was the result of two persons acquiring explosive devices, delivering them to specific locations, and setting off the attacks. In practice, one seeks the extent to which $\boxed{\text{PERSON}} \overset{acquire}{\rightarrow} \boxed{\text{DEVICE}} \overset{deliver}{\rightarrow} \boxed{\text{PLACE}}$ necessarily implies a $\overset{set-off}{\rightarrow} \boxed{\text{ATTACK}}$. Instead of prediction or causality, just as important is to demonstrate **association**, which is a looser concept and can be intuitively justified with the following:

1. **Event support:** An event does not happen at random. It requires prior support, whether financial, logistical, or others, which is described in Big Data, and can be clearly-worded or implicit. Mathematically, it can be stated that when n entities are observed in a spatial region, then there exists an entity $n + 1$ which is bound to be observed as well. This denotes *Bayesian Inference* and by extension *Distance-based Bayesian Inference*;
2. **Event influence:** Events may affect other events propagating through different regions. This means that an event in one area can influence a different event in a different area, allowing one to compute a *Spatial Association Index*, which is described later;
3. **Event interpretation:** Violent events unfold as a consequence of prior developments. While Big Data may describe these events in detail, it is often contaminated with noise that masks the true development. This leads to the notion of association in terms of *Spatio-logical Inference* in which a large number of possibilities that explain a violent event can be reduced to the most probable ones;

The above items ground relatedness between entities and events described in Big Data, giving rise to the association strategies that we propose and explain in the next subsections.

3.2 Analysis framework

At a high level, the work proposed in this study follows the steps shown in Fig. 3. Briefly, entities are extracted from the datasets (step ❶), geocoded and identified with timestamps (step ❷). Storylines are then generated (step ❸), and finally, association analysis is performed (step ❹). To generate the storylines, the work detailed in [24] is reused, and briefly summarized here.

That approach takes as input a dataset with entities for which locations and timestamps are available (or can be obtained). Locations are geocoded into latitudes and longitudes, and entities are extracted, stored, and indexed spatially. Relationships between entities are also extracted. A relationship is an interaction between two entities, such as when "person1 talks to person2", in which case the relationship is "talks to". An entity graph is then built by creating links among the extracted entities using the extracted relationships. For each entity in the graph, a *ConceptRank* (i.e., a variation of *PageRank*) is calculated. The storylines are formed in 3 steps: (1) the user selects an entity to be the entrypoint, i.e, the point from where the story begins; (2) from the entrypoint, the algorithm applies *Ripley's K function* to find an optimal radius within which the concentration of entities is dense; (3) within that radius, the entrypoint is linked to the top-k entities of highest *ConceptRank*, sorted in time order. The set of linked entities that this process generates is the final storyline, which has the general format $\boxed{\text{entity-1}} \overset{relationship-1}{\rightarrow} \boxed{\text{entity-2}} \overset{relationship-2}{\rightarrow} \boxed{\text{entity-3}}$. The length of the storyline may vary without bound.

The steps described above are suitable for Big Data for three reasons: (1) graphs provide a scalable data structure that can handle increasing numbers of entities and events; (2) ConceptRank can be computed in a distributed environment for large graphs, or locally for subsets of a graph; (3) Entity connectivity can be constrained to the ones relevant for particular domains, and disregarded for others.

Because the number of generated storylines can be massive (a consequence of Big Data), an intermediate step should be taken to perform hierarchical clustering. The clustering process serves two purposes: segregate the events in the storylines into related groups and allow processing to be done on a per-cluster basis, which is more manageable. In the final step,

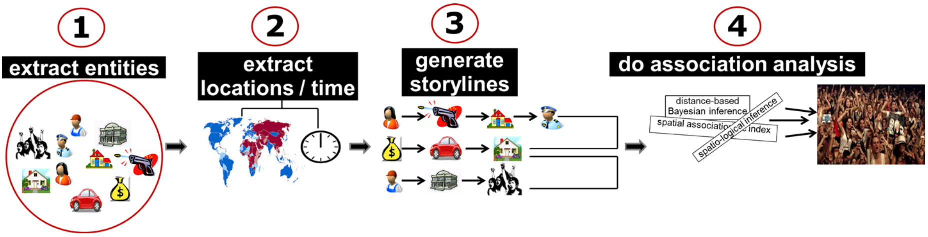


Fig. 3 Associative process using spatio-temporal storylines. In steps 1 and 2, entities are extracted from the input data sources along with their locations and timestamps. Storylines are subsequently generated from them (step 3). Further in step 4, the storylines are used as input to three event association approaches: *Distance-based Bayesian Inference* (*dbB*), *Spatial Association Index* (*SAI*), and *Spatio-logical Inference* (*SLI*). Each of these three methods respectively output a numerical association score: a probability of event occurrence, a measure of association between storylines, and a measure of compatibility between events

the three methods mentioned previously are explored to reason over violent events, providing intuitive justifications for their use. Those methods, *Distance-based Bayesian Inference* (*DbB*), *Spatial Association Index* (*SAI*), and *Spatio-logical Inference* (*SLI*) provide the foundation for Sections 3.3, 4.1, and 4.2. In the scope of this study, a storyline describes an event (or development) as the interaction among entities linked by relationships. A violent event is one that causes hardship at the individual, organizational, or governmental levels. Physical harm does not need to be involved. Unless otherwise stated, the following definitions will apply going forward:

Definition 1 An entity e represents a person, location, organization, event, or object described in a document. Only entities for which a location and a timestamp can be obtained are considered in this study.

Definition 2 A relationship, connection, or link defines a unit of interaction between two entities and is denoted by $e_i \xrightarrow{interaction} e_j$. All relationships $e_i \xrightarrow{interaction} e_j$ are intended to be directional.

Definition 3 A *trigger event* or a *final event* represents a real-world development extracted from text, such as an "explosion" or a "protest". They can be user-defined or application-specific based on an external ontology.

Definition 4 A storyline is a time-ordered sequence of n entities $\{e_1, \ldots, e_n\}$ where consecutive pairs (e_i, e_j) are linked by one relationship. The number of entities n is the length of the storyline.

3.3 Distance-based Bayesian inference

With traditional *Bayesian Inference*, probabilities are calculated by viewing each storyline as a *Bayesian Network*, in which each entity represents a node specified by a *Conditional Probability Distribution*(*CPD*). Specifically, if a storyline is described by three entities 🅐→🅑→🅒, one may want to find out its likelihood of occurring again, which is given by the joint probability of that entire storyline:

$$P(A, B, C) = P(A) \times P(B|A) \times P(C|B) \tag{1}$$

or, alternatively, one may want to simply find the probability of observing 🅒 knowing that 🅑 was observed in the past:

$$P(C|B) = \frac{P(B|C) \times P(C)}{P(B)} \tag{2}$$

Given the above, association can be determined either for single entities (or events) or for the entire storyline. In either case, the frequencies of all entities associated with a storyline must be known a priori. Figure 4a shows an entity graph related to the *Boston Marathon Bombings* of 2013 and Fig. 4b lists five of its possible storylines (S_1 through S_5). Assume those five storylines represent the entire available dataset. One piece of available prior knowledge is the killing of police officer $\boxed{\text{S.COLLIER}}$. Now assume one would like to know the likelihood that another police officer will

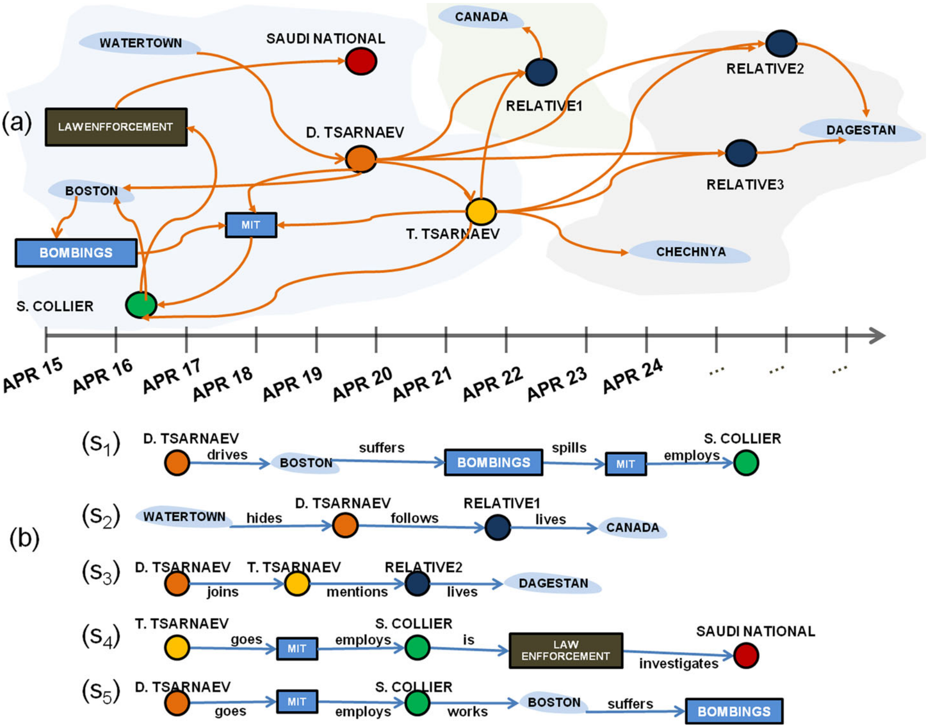

(a)
WATERTOWN
SAUDI NATIONAL
CANADA
RELATIVE2
LAWENFFORCEMENT
RELATIVE1
D. TSARNAEV
DAGESTAN
BOSTON
RELATIVE3
MIT
T. TSARNAEV
BOMBINGS
CHECHNYA
S. COLLIER
APR 15
APR 16
APR 17
APR 18
APR 19
APR 20
APR 21
APR 22
APR 23
APR 24
...
...
...
(b)
(S1) D. TSARNAEV drives BOSTON suffers BOMBINGS spills MIT employs S. COLLIER
(S2) WATERTOWN hides D. TSARNAEV follows RELATIVE1 lives CANADA
(S3) D. TSARNAEV joins T. TSARNAEV mentions RELATIVE2 lives DAGESTAN
(S4) T. TSARNAEV goes MIT employs S. COLLIER is LAW ENFFORCEMENT investigates SAUDI NATIONAL
(S5) D. TSARNAEV goes MIT employs S. COLLIER works BOSTON suffers BOMBINGS

Fig. 4 *Boston Marathon Bombings* spatio-temporal sequence. In **a**, each shape represents an entity observed in the data source. The edges denote relationships between the entities. In **b**, S_1 through S_5 represent five storylines connecting different entities. The English verbs define their relationships and correspond to the edges of the concept graph in (**a**)

be murdered in the near future. The best answer lies with S_4, which is the only storyline that contains a $\boxed{\text{LAW ENFORCEMENT}}$ presence and also someone related to a previous similar crime ($\boxed{\text{T.TSARNAEV}}$). Numerically, this likelihood corresponds to the joint probability of that storyline in relation to all the other four storylines: $P(\boxed{\text{T.TSARNAEV}}) \times P(\boxed{\text{MIT}} \| \boxed{\text{T.TSARNAEV}}) \times P(\boxed{\text{S.COLLIER}} \| \boxed{\text{MIT}}) \times P(\boxed{\text{LAW ENFORCEMENT}} \| \boxed{\text{S.COLLIER}}) = \frac{2}{5} \times \frac{1}{5} \times \frac{3}{5} \times \frac{1}{5} = 0.0096$. The following can then be stated: given this limited data, there is less than a 1 % chance of another police officer being murdered in the vicinity of the Boston area. This is traditional *Bayesian Inference*, which works well for highly-frequent storylines, but poses two problems for storytelling: entities must match perfectly (the ambiguity that comes with Big Data is a big problem!) and it does not consider the aspect of location. Next, we propose a method that relieves these two issues.

An intuitive approach to determine association is to simply search the data space for similar storylines that reoccur in constant time intervals. For instance, if $\boxed{\text{FLOOD}} \overset{causes}{\rightarrow} \boxed{\text{CHOLERA}} \overset{promotes}{\rightarrow} \boxed{\text{VIOLENCE}} \overset{affecting}{\rightarrow} \boxed{\text{SUDAN}}$ is observed every 5 years, then one can assume this pattern will be observed again in the next five-year interval.

In many applications, however, sequences such as those seldom repeat in a perfect manner. But they may reoccur with slight variations and in different places. Thus, flood may still be associated with violence, but perhaps not due to cholera, and maybe not in Sudan. Malaria may instead be the new factor in Zambia. This notion of spatial variability permits us to define ‘ontologically-similar storylines” in the following manner: (1) two storylines are ontologically similar if the location of at least one entity in one storyline is within a d distance of the location of an entity in the other storyline; (2) and apart from location, the two storylines must share at least one entity. And unlike what traditional *Bayesian Inference* would require, entities must not match perfectly. As long as the entities belong to the same ‘concept’ or ‘category’, they are deemed to be the same, such as “cholera” and “malaria”. Similarity, in this case, is determined by a user-defined ontological structure appropriate for a specific application domain.

In the above discussion, as long as two storylines are in close spatial proximity and share at least some characteristics, then they are deemed to be associated. For example, assume that for any given day, either storyline S_1 = 🅐→🅑→🅒→🅓 or storyline S_2 = 🅐→🅩→🅒→🅓 has appeared for the past year (for simplicity, letters are used for entity names and relationship tags are not shown). Assume also that 🅐 is the location on both S_1 and S_2. Since they have the same location, and share two other entities (🅒 and 🅓), then it can be stated that S_1 and S_2 are ontologically similar. This idea can be seen in Fig. 4b where the concept of *MIT* could be replaced with any “school” or “university”, and *S. Collier* could be regarded as any “law enforcement officer”, not just one specific person. Because Big Data may manifest itself in unlimited formats, we adopt this relaxed definition of similarity. Knowing that “cholera” and “malaria” can be treated as “diseases” saves many computing cycles, and allows greater data coverage than what an algorithm would normally accomplish under the magnitude of Big Data. And since these storylines may now be considered to be the “same”, traditional *Bayesian Inference* can be applied on them to find their probability of occurrence. This is what we define as *Distance-based Bayesian Inference*, which relaxes location and typing, and is attractive for its simplicity. As part of the experiments, we present this method as one of the association strategies.

4 On the association of violent events

Sections 4.1 and 4.2 present approaches to discover event associations based on spatial influence and event relatedness.

4.1 Spatial association index

A more powerful aspect of associating violent events, however, is to measure influence, i.e., whether the observation of a storyline in one place influences the occurrence of another storyline in another place. This is especially helpful under Big Data because events tend to be highly intermingled, and thus relating or pulling them apart becomes difficult. The *Boston Marathon Bombings*, for instance, provoked a myriad of reactions ranging from street closures around the blast site to a shootout in Watertown, a nearby area. In other words, an event in area A triggered other events in areas B, C, D, etc. At a high level, this is spatial correlation [28] framed in terms of entities and their interactions, rather than through traditional comparison of specific attributes, as in the work of [31].

The first consideration to be made is the following: if the influence of area A on area B is high, then there is a high likelihood that whenever A experiences a storyline, there exists

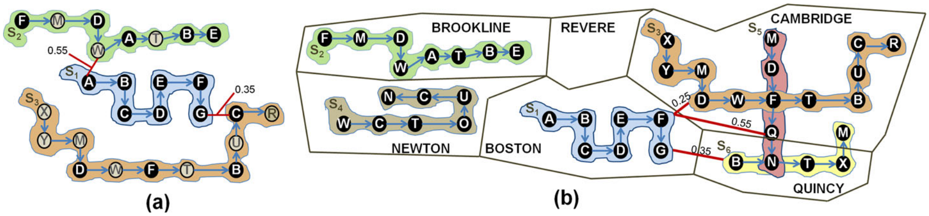


Fig. 5 Hypothetical set of storylines located in different regions. **a** Three storylines of different numbers of entities. **b** Six storylines spread across various cities. The *circles* denote entities and the edges represent relationships. *Red lines* denote the shortest normalized distances between corresponding storylines

other storyline(s) that B will experience (the analogous case for low values is the same). The goal then is to find out those storyline(s) that B will experience and identify the violent events behind them. Note that the storylines observed by A could, but need not be the same as the storylines observed by B. In order to gauge the level of influence between locations given their respective storylines, we propose a *Spatial Association Index* (*SAI*) below.

4.1.1 Comparing storylines

The influence between two entities is perceived to be stronger when they are located within a reasonably-short distance of one another, and thus location is an important aspect. In addition, for associations to happen, there must exist a minimum amount of commonality that bridges the two locations. In other words, events must not only be spatially close, but must also share entities. We combine these ideas to design the index below. First, the following definitions are necessary:

Definition 5 The distance between storyline S_x, composed of entities $\mathrm{E} = \{e_1, \ldots, e_n\}$ and location l_y, denoted *dist*(S_x,l_y), is the shortest distance between any $e_i \in \mathrm{E}$ and any point in l_y.

Definition 6 The distance between two storylines S_x and S_y, composed respectively of entities $\mathrm{E}_x = \{e_1, \ldots, e_n\}$ and $\mathrm{E}_y = \{e_1, \ldots, e_n\}$, and denoted *dist*($\mathrm{S}_x$,$\mathrm{S}_y$), is the shortest distance between any $e_i \in \mathrm{E}_x$ and $e_k \in \mathrm{E}_y$.

Def(s). 5 and 6 establish distance as a function of the closest entity to a specific location or to another entity in space. Here, distance is treated in spatial terms, and preferably using *metric* measures such as *Euclidean*, since they conform to symmetricity, which simplifies distance computations, and is highly desirable for Big Data. In practical use, however, other metrics can be just as applicable. Using the above definitions, we propose the *Spatial Association Index* (*SAI*) between two storylines S_x and S_y as:

$$SAI(S_x, S_y) = \log\left\{\frac{1}{dist(S_x, S_y)} \times n\right\} \tag{3}$$

where n is the number of shared entities between storylines S_x and S_y (if there is a high discrepancy between the number of entities between storylines, then normalizing n is necessary based on the shortest and longest storylines of the dataset). Equation 3 indicates that shorter distances and high numbers of shared entities contribute to a larger value, which indicates a stronger level of association, and is indeed the desired effect. As an example, Fig. 5a shows three storylines, S_1, S_2, and S_3, all of different lengths. S_1 and S_2 share 5 entities (Ⓐ,Ⓑ,Ⓓ, Ⓔ,, and Ⓕ). The two closest entities between S_1 and S_2 are Ⓐ and Ⓦ, which at a distance of 0.55 (normalized on a [0,1] scale), determine the distance between these two storylines. Calculating their *SAI*, therefore, yields $SAI(S_1, S_2) = \log\left\{\frac{1}{0.55} \times 5\right\} = 0.96$. Repeating the calculation for S_1 and S_3 results in $SAI(S_1, S_3) = \log\left\{\frac{1}{0.35} \times 4\right\} = 1.05$.

Comparing the two results, one can then claim that storyline S_1 is more tightly associated with storyline S_3 than to S_2. In everyday language, these results would be akin to stating that whenever events of the first storyline happen in one location, they are more likely to be followed by events of the third storyline. Note that the *SAI* values are not restricted to the range [0,1], and thus, are not probabilities. Rather, they are a spatial measure of influence that can be used to compare storylines. True probabilities can be computed using *Bayesian Inference* as described previously.

Equation 3 requires that two storylines be supplied ahead of time. In exploratory analysis, however, one may want to investigate not simply two storylines, but rather the influence of a source location on a target location based on their respective storylines. A classical example are protests, which many times originate peacefully in a small area and spread as looting, fights, and other acts of violence in various directions. In such a scenario, influence is better understood as a location-to-location process based on an initial random source storyline S, which is discussed next.

4.1.2 Comparing locations

For location to location, what is initially given is one storyline. Then the goal is to determine the influence of its location on other nearby locations. To achieve this redefined notion of influence, we reuse the *SAI* index above in the following algorithm:

1. **Identify locations**: starting from a user-specified source storyline S in a desired area of study, identify the closest location to S that shares at least one entity with S. Label the identified location L_{target} and the location of the closest entity in S as L_{source}.
2. **Retrieve storylines**: find the set of all storylines that refer to location L_{target}. Call that set *ALL-STORYLINES*.
3. **Calculate the index**: using *ALL-STORYLINES*, compute the *Spatial Association Index* of L_{source} on L_{target} w.r.t. S:

$$SAI(L_{source}, L_{target}, S) = \sum_{i=1}^{|ALL-STORYLINES|} SAI(S, ALL-STORYLINES_i) \tag{4}$$

The above algorithm can be used on Big Data in two ways. For textual characteristics of entities and events, a search engine based on TF-IDF and an *inverted file* index provides quick access to millions of records even when several ontological schemes are used. For locations, a spatial index, such as *R-tree*, permits efficient retrieval of spatial features at varying resolutions, such as per address, state/province, or country. The above algorithm operates in the following manner: given a source storyline, it finds the closest nearby region that also has storylines with similar entities (at least one). It then investigates all of the discovered storylines for that nearby region, calculating their *SAI* values, and summing them up into one aggregated value. This aggregated value represents a numerical measure of storyline influence between the originating location (source) and the investigated location (target). Again, the higher the *SAI*, the stronger the level of association between the locations. A visual example follows.

Figure 5b shows five regions (*Brookline*, *Newton*, *Revere*, *Cambridge*, and *Quincy*) around the *Boston* area. Except for *Revere*, all areas contain at least one storyline, and some of their entities are shared across regions. This is the case of entity 🅕, which is observed at different times in *Brookline*, *Boston*, and *Cambridge*. Imagine that an analyst would like to understand how the events in *Boston* imply events in those other areas. Following the algorithm above, the analyst would first identify the closest area to *Boston* that has a storyline which shares one or more entities with a *Boston* storyline. It turns out that storylines of all areas share entities with the *Boston* storylines. In this case, the chosen location is *Cambridge* since it is the closest to *Boston* considering driving distance (when several locations are equally distant, the one with the highest number of common entities is selected before a random choice is made). Thus, according to *step 1*, $L_{source} = Boston$ and $L_{destination} = Cambridge$. As per *step 2*, we now retrieve all storylines associated with *Cambridge*, which according to Fig. 5b are S_3, S_5, and S_6. In the last step, we compute all *SAI* values between S_1 and each of S_3, S_5, and S_6 (Eq. 3), and sum them up (Eq. 4). Considering the storyline distances shown by red lines in Fig. 5(b), the computations would be: $SAI(Boston, Cambridge) = SAI(S_1, S_3) + SAI(S_1, S_5) + SAI(S_1, S_6) = \log\left\{\frac{1}{0.25} \times 4\right\} + \log\left\{\frac{1}{0.55} \times 2\right\} + \log\left\{\frac{1}{0.35} \times 1\right\} = 2.21$. One could certainly perform the same calculations for any other areas, e.g., *SAI(Boston,Brookline)*, and compare their *Spatial Association Index*.

The algorithm outputs one *SAI* value for each pair of storylines. Optimizations can be done, such as pruning locations known to be uninteresting, or removing storylines known to be uninformative. Intuitively, this approach allows the analyst to see how events propagate in time and space. It does so by providing a numerical value of confidence that developments in one place will be followed by developments in another place. It should be noted, however, that this does not translate to a prediction. In prediction, one fact implies that another fact will occur, having a strong implication to cause and effect. Here, we are simply stating that two or more facts will be observed in sequence, without any assumption that one will cause the others. This is the motivation as to why the *SAI* has a strong association potential, and thus its name. For example, one could state that a “bombing in *Boston*” is strongly associated with “law enforcement in *Cambridge*” with *SAI* x. To avoid specific scenarios unlikely to repeat (such as the *Boston Marathon Bombings*), a better assertion is that $action_1$ in location A is associated with $action_2$ in location B, when dist(A,B)$\leq$ distance d and their *SAI*$\leq$ threshold t.

Later in this paper, the experiments will demonstrate how storylines observed in certain locations are associated with seemingly disparate events in other locations. These experiments use real datasets related to social unrest in Mexico.

4.2 Spatio-logical inference

As mentioned earlier, violent events can be viewed as the end result of larger processes composed of one or more *trigger events*. In the *Poll Tax Riots*, for example, some of those *trigger events* were identified, two of which were that activists organized protests and police closed some streets. Intuitively, each of these *trigger events* contribute a certain amount of momentum to the riots, with some weighing in more heavily than others. The goal here is

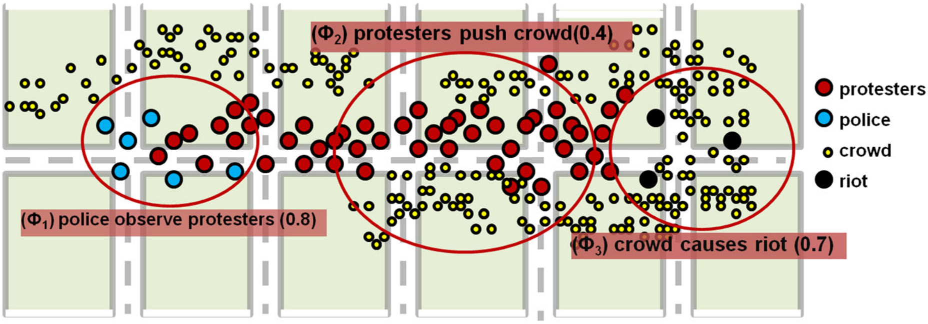

(Φ2) protesters push crowd(0.4)
(Φ1) police observe protesters (0.8)
(Φ3) crowd causes riot (0.7)
protesters
police
crowd
riot

Fig. 6 A spatial diagram of entity interactions enclosed in ovals. The l*eft and center ovals* represent *trigger events*, and the rightmost one is the *final event*. Each event has a text description, is denoted by ϕ_1, ϕ_2, and ϕ_3, and has a *soft truth* value. The sequence conveys a storyline in which as police observe protesters, and protesters push against the crowd, a riot ensues

to come up with these weights, which are called "*soft truths*", such that, when put together, the final violent event can be deemed probable or not.

A *soft truth* is simply a *numerical belief* in the range [0,1] that two entities will interact in a particular way. Thus, one person may have seen police observing protesters and assign this fact a *soft truth* of 0.75 (almost certain). Another person, on the other hand, was not sure the police was involved, lowering the *soft truth* to 0.25 (not certain). Under this approach, Big Data can be viewed as a vast collection of soft truths that can be manipulated in a piecewise fashion. The combination of event sequences and *soft truths* allows one to generate rules and determine how well they lead to the violent event (i.e., their *distance to satisfaction*), which is explained below.

4.2.1 Rule processing

Informally, this problem can be expressed as follows: given a storyline composed of several interacting entities which leads to a final violent event, a method is needed to combine the individual *soft truths* of each interaction. We can then use the the prior events to make a decision of whether the consolidated interactions are compatible with the final violent event or not. If they can generate the violent event, then we say that the prior events are compatible with the final event.

Consider Fig. 6 which depicts different sets of entities (*police*, *protesters*, *crowd*) interacting among themselves in the streets. There are three interactions, denoted ϕ_1, ϕ_2, and ϕ_3, each described in text with an associated *soft truth* value. The *soft truths* can be obtained from various sources: historical frequencies, input of domain experts, and random sampling, among others. An algorithmic method is needed to answer the following question: is the combination of "*police observe protesters*" (ϕ_1) and "*protesters push against crowd*" (ϕ_2) enough for the crowd to "*cause a riot*" (ϕ_3)? Formally, this problem can be modeled in *First Order Logic* with the following statement:

$$\textbf{observe(police,protesters)} \wedge \textbf{push(protesters,crowd)} \implies \textbf{cause(crowd,riot)} \qquad \text{(Rule r}_1\text{)}$$

The above statement establishes a logical rule (r_1) that relates two *trigger events* via an "and" relationship ($\wedge$) to the *final event*, which is the riot. All of these events are in the format ***predicate***(***entity***$_x$,***entity***$_z$). It should read that *entity*$_x$ performs the *predicate* on *entity*$_z$, meaning that when police observe protesters and protesters push against the crowd, it implies that a riot will break out. This type of statement represents hard logic, i.e., it determines whether developments will or will not happen, such as in a binary fashion. In terms of violent events, hard logic in many instances is not applicable because one can seldom state with certainty that a riot will or will not occur. For this reason, instead of hard logic, a more appropriate way of reasoning over this type of question is to relax the binary restriction, and permit interactions to have a *soft truth* in a continuous fashion. Relaxing these restrictions allows Rule r$_1$ to be rewritten as follows:

0.25: **observe(police,protesters)**(0.8) $\wedge$ **push(protesters,crowd)**(0.4) $\Longrightarrow$ **cause(crowd,riot)** (0.7) (Rule r$_2$)

0.44: **observe(police,protesters)**(0.9) $\wedge$ **push(protesters,crowd)**(0.3) $\Longrightarrow$ **cause(crowd,riot)** (0.1) (Rule r$_3$)

And generalizing them:

$$\text{RW} : \phi_1(\mathbf{e}_a, \mathbf{e}_b)(w_1) \wedge \ldots \wedge \phi_n(\mathbf{e}_u, \mathbf{e}_v)(w_n) \Longrightarrow \phi_{n+1}(\mathbf{e}_w, \mathbf{e}_z)(w_{n+1})$$

where RW is the rule weight, ϕ_i is either a *trigger event* or the *final event*, e$_i$ represents an entity (or set of) and w$_i$ is a *soft truth* value. Note that *trigger events* always appear in the antecedent of the rule (i.e., before the $\Longrightarrow$ sign), and the *final event* always appear in the consequent of the rule (i.e., after the $\Longrightarrow$ sign). In Rules r$_2$ and r$_3$ respectively, the *trigger events* have *soft truths* (0.8, 0.4, 0.9, 0.3) and the *final events* have *soft truths* (0.7, 0.1). The rules themselves have weights 0.25 and 0.44. In practice, the rules put in formal notation statements related to what "people think" or "may have seen" or "has happened" given uncertainty. There could be different rules that also lead to the same outcome (i.e., the riot), such as:

0.65: **seen_with(weapons,protesters)**(0.8) $\wedge$ **push(protesters,crowd)**(0.4) $\Longrightarrow$ **cause(crowd,riot)** (0.7) (Rule r$_4$)

Given its higher rule weight, Rule r$_4$ is preferable to r$_2$ and r$_3$ (possibly because it involves weapons!). In a real application, thousands of such rules can be generated, which should be expected in a Big Data scenario, and requires a numerical method to determine how good each rule actually is. In practice, we must find whether the *trigger events* satisfy the riot, and if not, their *distance from satisfaction*. What was described so far is derived from *Probabilistic Soft Logic* (PSL) [11]. PSL allows one to determine if the *trigger events* of a rule satisfy the *final event* for that same rule. If they do, one can then state that the rule is compatible with the *final event*.

Given a set of *trigger events* $\phi = \{\phi_1, \ldots, \phi_n\}$, the assignment of $\phi_i \rightarrow [0,1]^n$ represents the allocation of a *soft truth* value to an interaction between two entities. This allocation is called an *interpretation* $I(\phi_i)$. PSL uses the *Lukasiewicz t-norm* and *co-norm* to relax the traditional logical conjunction ($\wedge$) and disjunction ($\vee$) into continuous values as follows:

$$I = \begin{cases} \phi_1 \tilde{\wedge} \phi_2 = max\{0, I(\phi_1) + I(\phi_2) - 1\} \\ \phi_1 \tilde{\vee} \phi_2 = min\{I(\phi_1) + I(\phi_2), 1\} \\ \tilde{\neg} = 1 - I(l_1) \end{cases} \tag{5}$$

The ˜ symbol is applied to denote the relaxed version of the normal logical operators, which allows us to assert the following:

Definition 7 Given a rule r, composed of a set of *trigger events* $\Phi = \{\phi_1, \ldots, \phi_n\}$ and a *final event* ϕ_{final} where each ϕ_i and ϕ_{final} have an interpretation in [0,1], r is satisfied if and only if $I(\phi_1, \ldots, \phi_n) \leq I(\phi_{final})$.

Definition 7 states that the interaction established by the entities in the *final event* (ϕ_{final}) must have at least the same *soft truths* as the interactions of its constituent *trigger events* ($\phi_1, \ldots, \phi_n$). The rule's distance to satisfaction for interpretation I is given by:

$$d_r(I) = max\{0, I(\phi_1, \ldots, \phi_n) - I(\phi_{final})\} \tag{6}$$

As an example, take Rule r_2, for which we wish to compute its distance to satisfaction $d_r(I)$. I(ϕ_1, ϕ_2) $=$ max$\{0, 0.8 + 0.4 - 1\} = 0.2$. Since 0.2$\leq$0.7, we say that the rule is satisfied and $d_r(I) = 0$. This contrasts with Rule r_3. where I(ϕ_1, ϕ_2) $=$ max$\{0, 0.9 + 0.3 - 1\} = 0.2$, and $d_r(I) =$ max$\{0, 0.2 - 0.1\} = 0.1$. Rule 3 is more distant to satisfaction than Rule 2. Interpretations can be challenging to deal with because different people have different opinions and different perceptions of facts. This method provides a way to show that some rules are more feasible than others from a numerical perspective. We now propose an algorithm that generate these rules.

4.2.2 Rule generation

In this section, we propose an algorithm to generate rules using spatial distance as one of their components, and thus the name *spatio-logical inference*. Look ahead to Table 6 and the discussion in Section 5.3 for a brief visual example. More formally, this process obeys the steps of Algorithm 1, explained below.

```
ALGORITHM 1 Candidate Rule Generation
inputs : set of $STORYLINES = \{s_1,...,s_n\}$ where each $s_i$ is composed of events $\phi_1, ..., \phi_m$ tagged by
         locations and timestamps in an area of study, number of desired rules $n$, size of rule $s$, distance
         $d$, event-pair Probability-Matrix
output: set of weight-based rules $RULES$
Initialize
$|Rules| = 0$; $\phi_{final} \leftarrow \phi_k$ ; // select one event in the dataset to be the final event
Pre-processing Stage
while STORYLINES exist do
    foreach pair $(\phi_i,\phi_j) \in \{s_i\}$ where $\phi_i \neq \phi_j$ do
        Distance-Matrix $\leftarrow$ store(normalizedDistance($\phi_i$,$\phi_j$)) ; // calculate the distance between each pair of entities.
    end
end
Main Stage
while $|Rules| \leq n$ do
    List{Trigger-Events} = query(Distance-Matrix,$\phi_{final}$,$s$,$d$) ; // perform a query for the $s$ closest events within distance $d$ of
    the final event.
    rule $\leftarrow$ concatenate(List{Trigger-Events},"$\wedge$",$\phi_{final}$) ; // combine all trigger events to the final event with an "and"
    relationship.
    foreach $(\phi_i,\phi_j) \in rule$, $\phi_i \neq \phi_j$, do
        set soft-truth($\phi_i$,$\phi_j$) = Probability-Matrix[($\phi_i$,$\phi_j$)] ; // set the soft truth for each interaction in the rule by looking up the
        probability of its composing events in the probability matrix.
    end
    set $rule_{RW} = \frac{1}{avgDistEvents(rule, Distance-Matrix)}$ ; // set the rule's weight as the inverse of the average normalized distance among all its
    composing events
    $RULES \leftarrow$ rule ; // store the formed rule.
    increment $d$ ; // increase the search distance and perform another query.
end
output $RULES$;
```

The algorithm takes as input a set of storylines composed of many events. Each event is associated with a location (latitude and longitude). Because Big Data have the potential to explode the number of locations, the analysis should be constrained to refined regions where most entities reside or where most events take place. Several regions can be investigated at a time. To alleviate the high variability of Big Data, an ontological scheme should coalesce the most common concepts into simpler categories. The user must input the following items: the number of desired rules to be generated (n), the desired size of each rule (s), and a matrix of probabilities where each cell contains the likelihood of observing the corresponding events (event pair *Probability Matrix*). This matrix is obtained from historical data (in the experiments, it uses the *GDELT* dataset). Rule size is defined as the number of *trigger events* that composes the rule, i.e., the number of events concatenated by the $\wedge$ relationship. In the previous example, the size of Rule 4, for instance, is 2. The algorithm first initializes two items: *RULES*, a data structure to hold the final rules, as empty; and the user-selected *final event* ϕ_k (line 1).

Pre-processing stage After initializing the final event to be targeted, the algorithm first computes the distance between all events in the area of study, shown in line 3, to be used later. The results are stored in a *Distance-Matrix* (line 4).

Main stage First, using the *Distance-Matrix*, a query finds a number s of events (i.e., a number that matches the rule size) within a user-specified spatial distance d of the *final event*. The results are stored in List{Trigger-Events} (line 8). The rule is then formed by concatenating the found *trigger events* in the list to the *final event* ϕ_{final} via the “and ” ($\wedge$) operator (line 9). What remains to be done is to set the *soft truths* for each event in the rule. This is represented in lines 10 and 11 by doing a lookup in the probability matrix already provided. The overall rule weight is obtained by averaging the distances of all events for that rule, which can be obtained from the Distance-Matrix (line 13). The formed rule is then stored in the output data structure *RULES* (line 14) and the distance is incremented for a new search for more *trigger events* (line 15). The process continues until the desired number of rules has been reached, at which point the *RULES* are output in line 17. In terms of computational complexity, Algorithm 1 has its costliest step in building the *Distance-Matrix*. Since every pair of events must be compared, that step operates in O(n^2). However, this is in reality less of a problem since it is a one-time operation. In addition, since the identified locations tend to repeat, preprocessing them again is not necessary. Building the list of *Trigger Events* operates in O(n), since it comprises matching one specific event with a list of other events. Setting the soft truths for each event represents a look-up in the *Probability-Matrix*, which runs in O(1) (constant time). Therefore, in the worst case, the algorithm runs in O(n^2).

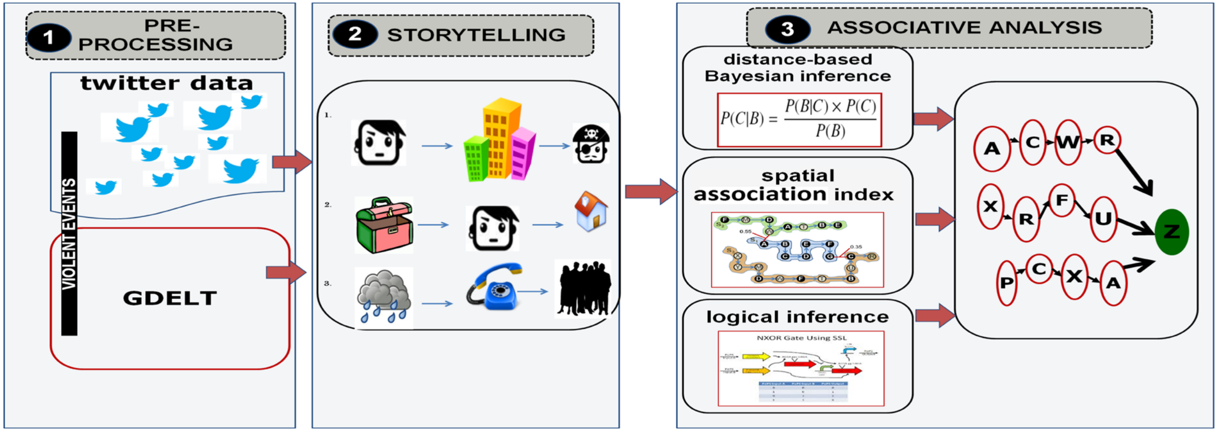
5 Empirical evaluation and technical discussion

The goal of the experiments is to investigate how the three methods described in Sections 3 and 4 can be employed to reason over real-world developments described in Big Data. Section 5.1 describes the specifications of the experiments. Section 5.2 evaluates *SAI* in two different modes. The *SLI* method is discussed in Section 5.3. Section 5.4 contrasts the three different methods, with key observations given in Section 5.5.

5.1 Experiment setup

The experiments follow the specifications of Table 1 and the steps in the associated image. Initially, *Twitter* and *GDELT* data related to violent events are ingested. As part of pre-processing, necessary clean-up steps, such as removing retweets, are performed. The storytelling process follows the approach in [24], and is briefly explained in Section 3.2. It

Table 1 Methodology and data specification of the experiments



Comparative methods	Number of records	Source	Years	Nature of violent events	Measure	Validation set
SAI	9,800,000	*Twitter*	2011, 2012, 2013	education reform in Mexico	recall-1	*GSR*
SAI	5,156,000	*GDELT*	2011	war (Afghanistan and Middle East)	recall-2	*GDELT*
SLI	3,200,000	*GDELT*	2011	war (Afghanistan and Middle East)	precision-1/recall-3	*GDELT*
DbB, *SAI*, *SLI*	2,580,000	*GDELT*	2011	war (Afghanistan and Middle East)	precision-1/recall-3	*GDELT*

The image shows the three-step association analysis using spatio-temporal storytelling on *Twitter* and *GDELT* data. The pre-processing stage performs data collection and clean up. Stage 2 generates storylines from the ingested data. In the associative analysis stage, the generated storylines are used as input to the three proposed methods (*Distance-based Bayesian Inference*, *Spatial Association Index*, and *Spatio-logical Inference*) with which association scores are computed

DbB=Distance-based Bayesian Inference, *SAI*= Spatial Association Index, *Li*= Spatio-logical Inference extracts entities and their relationships, geocodes them, identifies timeframes, and links the entities into storylines. In the associative analysis stage, the generated storylines are fed to the proposed methods from where different association scores are computed according to each method.

Data specification Two data sources were utilized: tweets spanning the years of 2011, 2012, and 2013; and *GDELT* data from 2011. Several experiments were performed with a varying number of records used in each, as shown in Table 1. The data contained a high variation of content: violent events reported in tweets and *GDELT* interactions, events of a non-violent nature, and a large number of other records without any apparent association. Two event types were targeted: education-related protests in *Mexico* and wars in *Asia*.

Comparative methods Evaluation was done on the three methods proposed in Sections 3 and 4 (*Distance-based Bayesian Inference*(*DbB*), *Spatial Association Index* (*SAI*), *Spatio-logical Inference* (*SLI*)), which are designated as the 'comparative methods'). Two directions were taken: first, they were investigated and discussed separately, and second, they were compared to one another. The following approach was taken: a subset of the data was applied to the comparative methods to see what they were able to associate, and then those findings were validated in a different subset of the data not used previously. This is somewhat akin to a train-and-test approach. The number of records for each subset is specified in the experiment sections where they are discussed. In trying to approximate the high variability of Big Data, the first examined method, *SAI*, used both *Twitter* and *GDELT* data employing different record sets as shown on the table. *SLI* was also investigated separately using 3,200,000 *GDELT* records. For the part of the experiments that compared all methods, 2,580,000 *GDELT* records were utilized. Those records were different from the previous ones. No specific data distribution was assumed, but areas of study where violent events were known to be of a high enough frequency were used, such that associative analysis was actually plausible.

Performance measures The evaluation's purpose was to provide a variation of discussions, and thus different directions were taken. The main goal was effectiveness, i.e., finding relevant stories among millions of possibilities. In some of the experiments, only visualization of the results was performed with the intent of providing an intuitive perspective. For individual analysis, **recall** was selected as the performance measurement, leaving out **precision** for simplicity. Yet, when all comparative methods were compared, both **precision** and **recall** were used. In terms of *storytelling*, it must be noted that both precision and recall should be interpreted carefully, as different viewpoints may arise based on what one would consider a "true" association as opposed to a "missed" association. Likewise, storylines do not possess standard definitions for what should be considered "relevant" or "similar". For this reason, Table 2 defines this study's usage of precision, recall, "successful association", "relevant event", "similar events", and "unobserved events". As much as possible, the goal was to reflect the definitions of precision and recall according to traditional *Information Retrieval*, and provide a clear picture of what was being measured.

5.2 Association analysis using *SAI*

Unlike traditional probability, in which a score of 1.0 indicates full certainty, the *Spatial Association Index* (*SAI*) is not constrained by an upper bound. As a result, one *SAI* value on its own has little meaning. To be useful, it must be compared to or contrasted with other

Table 2 Explanation of performance measures

Measure	Meaning
recall-1= $\frac{identified-as-relevant}{(identified-as-relevant)+(relevant-but-not-identified)}$	fraction of events correctly identified as relevant over the set of all relevant events.
recall-2= $\frac{\vert SAI(unobservedevents)\geq\vert x\vert}{\vert unobservedevents\vert}$	fraction of the *unobserved events* that have an *SFI* value of x or more.
recall-3 = $\frac{similaridentified}{similaridentified+similarmissed}$	fraction of similar events that were identified over the total number of similar events.
precision-1= $\frac{similaridentified}{allretrieved}$	fraction of similar events that were identified over all retrieved records.

Definitions

Successful association: An initial storyline is successfully associated with a target storyline (i.e., association is deemed relevant) if the initial storyline is linked to the target storyline through at least one common entity and the target storyline has an event identified in *GSR* or in *GDELT*.

Relevant event: An event is deemed relevant if the storyline it belongs to makes a successful association of a target storyline.

Similar events: Two events are deemed similar if they both belong in the same ontological branch (according to a user-specified ontology) and are located within *distance to satisfaction* $\leq t$ of one another, where t is a user-defined threshold.

Unobserved event: An event is considered unobserved if it is not included in previous analysis.

SAI values. In a perfect world, higher *SAI* indices always indicate that a violent event will occur with higher likelihood than lower *SAI* indices. Thus, the highest possible *SAI* values are sought, and investigated if they translate to sets of associated events known to be true.

For this part of the experiments, the dataset was composed of 9,800,000 tweets related to *civil unrest*[1] in *Mexico* for parts of 2011, 2012, and 2013. Using these tweets, storylines were generated based on the approach described in [24]. Events related to *education reform* were targeted, which provoked social strife in *Mexico*, and were documented as part of the *Gold Standard Report* (GSR) from the *Intelligence Advanced Research Projects Activity* (IARPA) [10]. GSR served as the ground truth for this part of the experiments.

Two sets of experiments were performed: **(1) point-to-point mode**: pairs of initial and target storylines were investigated, and for each pair, their *SAI* values were calculated. For the top-*k* pairs of highest *SAI* values, corresponding events in GSR were identified. To determine whether the association was successful: if the target storyline was linked to the GSR event through at least one entity and had the same distance (or less) to the initial storyline, then the association was successful based on that *SAI* score; and **(2) point-to-region mode**: starting from a set of initial storylines, the *SAI* values to storylines of nearby regions were calculated. Then, those *SAI* values were compared between the different regions, finding matching events for the different regions in GSR, and justifying which ones were correctly associated (or not) as before. Each task involved four steps: **(a)** select an initial storyline; **(b)** calculate the *SAI* values between the initial storyline and the other storylines; **(c)** select a number of top-k *SAI* values; and **(d)** verify if those locations were the place of a violent event that is documented in the GSR list. In other words, they sought the regions whose *SAI* values translated to good **recall-1**$= \frac{identified-as-relevant}{identified-as-relevant+relevant-but-not-identified}$. This formula defined the fraction of events correctly identified as relevant over the set of events that should have been identified as relevant, but were not.

Table 3 illustrates on the top row an initial storyline, which is denoted S$_i$. This storyline, which was observed in *Mexico City* in January 2013, reports a teachers' strike for better financial conditions. Each row of the table shows a target storyline (S$_1$ through S$_{10}$) generated from tweets, the target storyline's location, its distance to the location of the initial storyline (*Mexico City*), the *SFI* value between the initial storyline and the target storyline, and a GSR event that confirms the veracity of the target storyline.

5.2.1 SAI in point-to-point mode

This subsection investigates association based on the locations of two specific storylines at a time, thus the "point-to-point" designation. Table 3 is sorted in decreasing *SAI* values. Immediately, it can be seen that the lowest *SAI* value tied to a successful association is 0.48 (the last row in the table). Since S$_1$ had the highest *SAI* value in the table, the first conclusion was that S$_i$ was associated with the target storyline S$_1$ better than it was associated with any of the other nine storylines. In other words, a **STRIKE** by the **TEACHERS** for better **SALARY** and **FUNDS** was deemed a strong indicator of **EDUCATION**-related fighting by the **SNTE** (workers' union) for better **SALARY** and **FUNDS**, which is documented in the corresponding GSR event. Note that both S$_i$ and S$_1$ have the same location (*Mexico City*), with zero distance of each other, which boosted their *SAI* value according to Eq. 3. They also shared most entities, shown in uppercase letters.

[1] civil unrest denotes an event of social impact, such as a strike or a protest.

Discussion At a distance of 28 km, S_2 was only somewhat farther from *Mexico City*, but had a much lower *SAI* score (1.52) than S_1. The lower score was due to two reasons: the longer distance between *San Pedro Atlixco* and *Mexico City*, and the fact that S_i and S_2 only share two entities (TEACHERS-TEACHERS and PROTEST-STRIKE). One notable item is S_{10}, whose storyline had the lowest *SAI* value of all (0.48), even though its distance to S_i (297 km) was much shorter than S_6, S_7, S_8, and S_9. It indicates that location is not the only determining factor in an associative strategy, though an important one. Looking at the table, it is generally true that longer distances determine lower *SAI* values, which should be expected. However, this assumption breaks in S_4 and S_5, which seem to hold a contradiction. The former is located farther away from S_i than the latter, but has a higher *SAI* value. The difference, again, is due to the number of shared entities with S_i, which is higher for the former than for the latter.

Based solely on this dataset, the premise is that, as a violent event, the **STRIKE** in *Mexico City* described in S_i is more likely to be followed by fighting by the **SNTE** also in *Mexico City* (S_1) than by a march by the **SNTE** in *San Pedro Atlixco* (S_2). It could be further stated that a **PROTEST** by **TEACHERS** in *San Pedro Atlixco* (S_2) was more probable than a **TEACHER**'s march against lower **BUDGET** in *Tlaxcala* (S_3). Such observations can be generalized into an associative model of how organizations mobilize people in social settings, which can be further applied in tasks such as classification or rule association mining.

Note that the above statements do not come solely from the comparison of a few storylines. Rather, they compare storylines that represent millions of entities involved in thousands of violent events. These results are successful because all of the target storylines were highly reflective of a real event (documented in the GSR), which is shown in the last column of the table.

5.2.2 SAI in point-to-region mode

The discussion now switches to point-to-region mode, in which the objective is to investigate the *SAI* values from an initial storyline to all other storylines contained in a different region. Thus, given the same initial storyline as in the previous example, the goal is to know if the **STRIKE** from the **TEACHERS** for better **SALARY** and **FUNDS** in *Mexico City* propagated to other regions as similar events, or even caused different events to happen. The higher the *SAI* value for a region, the higher the belief that storylines in that region would reoccur.

Discussion Some of the results are shown in Table 4. The first thing to notice is that L_1, in the vicinity of *Mexico City*, had 545 storylines that drove the highest average *SAI* value in the set (2.71). Noting that L_i and L_1 had the same location (*Mexico City*) and thus no distance between them, their high *SAI* value is not surprising. In practice, it would be similar to stating that violent events often spread to nearby areas, such as rioting along connected streets. A more interesting case is L_2, which contained a significantly smaller number of storylines (275), but not a much lower *SAI* value than L_1 (2.31). Two reasons explain this difference: first, *Pachuca* is not very far from *Mexico City* (87 km); second, Pachuca's storylines have a high average number of shared entities with S_1 (2.4). They helped boost the *SAI* value calculated with Eq. 4. *Veracruz*, in L_9, had a high number of storylines related to *education reform*, but its long distance to *Mexico City* (313 km) and a low number of shared entities with S_1 (1.5) gave it a low *SAI* score (0.54), making it challenging to associate events in *Mexico City* with any of *Veracruz*'s events (Fig. 7).

Table 3 Demonstration of the *Spatial Association Index* in point-to-point mode between an initial storyline and 10 target storylines

Initial storyline S_i: **STRIKE** affect **TEACHERS** demand **SALARY** higher **FUNDS**.				**Location:** *Mexico City*	
	Target storyline	**Location**	**Distance to Mexico City (km)**	**SAI**	**GSR Event**
S_1	**EDUCATION** fighting **SNTE** paying **SALARY** lower **FUNDS**.	Mexico City	0	3.60	SNTE Protesters block Eje Central; demand pension pay.
S_2	SNTE march **TEACHERS** participate **PROTEST**.	San Pedro Atlixco	28	1.52	SNTE teachers march in Atlixco.
S_3	**EDUCATION** march **TEACHER** lower **BUDGET**.	Tlaxcala	113	1.07	Teachers march against labor reform in Tlaxcala.
S_4	ROAD blocked **PROTEST** include **TEACHERS** ask **FUNDS**	Zitacuaro	129	1.02	Teachers block Morelia-Toluca in Zitacuaro.
S_5	**FIGHT** breaks **CITY** drain **FUNDS**.	Pachuca	87	1.01	Several incidents reported during SNTE's march.
S_6	**TEACHERS** lose **FUNDS** remove **BUDGET** impact **EDUCATION**.	Veracruz	313	0.76	SNTE teachers walk in Veracruz against education reform.
S_7	**EDUCATION** halt **UNIVERSITY** remove **STUDENT**.	Oaxaca	365	0.56	Stop at Oaxaca University affect more than 20 thousand students.
S_8	**TEACHER** protest **EDUCATION** lower **FUNDS**.	Aguascalientes	425	0.50	SNTE professors at Aguascalientes will march against education reform.
S_9	**FIGHT** break **STUDENT** distribute **FUNDS** sending **MORELIA**.	Michoacan	439	0.49	Teachers protest in Michoacan; demand Christmas pay.
S_{10}	**TEACHERS** march **CITY** protest **EDUCATION**.	Acapulco	297	0.48	In Acapulco, SNTE teachers from San Marcos will march.

The initial storyline is displayed across the top row. The storylines are related to *education reform* in *Mexico* generated from a set of approximately 9.8 million tweets. Each storyline is composed of entities (uppercase words) linked by relationships (lowercase words). Similar colors denote similar ontological concepts. Each target storyline has a location where it was observed, its distance to the initial storyline (*Mexico City*), and the *SAI* score w.r.t the initial storyline. The GSR event represents a development reported in the media that reflects the target storyline

Table 4 Demonstration of the *Spatial Association Index* in point-to-region mode between an initial storyline based out of Mexico City and 10 target locations

Initial storyline S_i :**STRIKE** affect **TEACHERS** demand **SALARY** higher **FUNDS**.					**Location L_i : Mexico City**
	Target location	Number of storylines in target location	Avg. number of shared entities	Avg. SAI	GSR Event
L_1	Mexico City	545	1.4	2.71	With protest, SNTE initiates informative campaign about education reform.
L_2	Pachuca	275	2.4	2.31	SME and CNTE protest in front of the Government.
L_3	San Pedro Atlixco	601	2.5	1.79	Protest takes place in front of Sagarpa's building.
L_4	Tlaxcala	325	2.0	1.26	CNTE teachers protest for eight hours, Metrobus service altered.
L_5	Zitacuaro	291	1.3	0.98	Teachers meet in front of the nation's Supreme Court.
L_6	Acapulco	255	1.2	0.87	Strike to continue at Autonoma University.
L_7	Oaxaca	184	1.2	0.75	DF Teachers will stop city center on Monday.
L_8	Michoacan	98	1.2	0.69	Teachers maintain pay dispute despite police confrontation.
L_9	Veracruz	402	1.5	0.54	Strike breaks out at Conalep plant in DF.
L_{10}	Aguascalientes	127	1.8	0.44	CNTE marches from Zacatecas to San Lazaro, maintain ground.

The initial storyline (S_i) and its location (L_i) are displayed across the top row. For each target location, the table shows the number of observed storylines, its average number of shared entities with the initial storyline, the average SAI values between the initial storyline and each storyline in that location, and a sample GSR Event in that target location

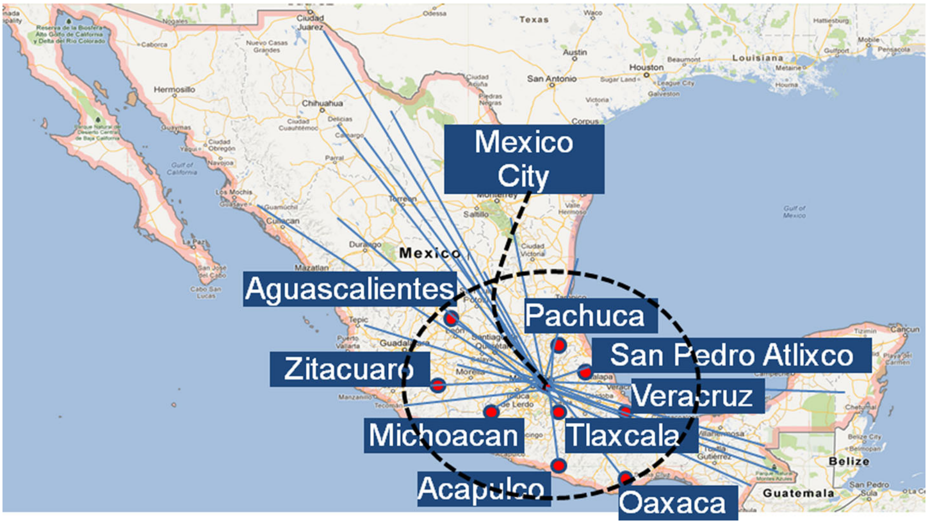


Fig. 7 Spatial propagation of *education reform* protests. Starting from Mexico City, similar events are observed around the country. The map shows 10 of approximately 1,000 affected locations

The importance of the spatial aspect of this study must be emphasized, showing that all items from L_1 to L_{10} were highly dependent on location. Without the spatial consideration, finding such events in Big Data would be unsurmountable. In this dataset, many of the storylines had no location explicitly stated. However, their related tweets did contain at least one metadata location that matched the location of the *GSR event*, and a timestamp that closely pre-dated the report of the event. This is particularly interesting in the case of L_{10}, whose *GSR event* was shown in *Zacatecas*, but whose *target location* was shown in *Aguascalientes*, which were only 43 km apart. The prominence of these storylines in close proximity of one another was significant for a simple reason: it indicated that the proposed *SAI* model based on spatial distance and shared entities could uncover related violent events that could reoccur in nearby areas in the future. If an analyst were interested in at most three regions of interest, Table 4 would allow her to speculate that from the initial storyline S_1, violent events with an *education reform* theme were more likely to take place in *Mexico City*, *Pachuca*, and *San Pedro Atlixco*, with decreasing order of confidence. The analyst might want to prioritize those regions.

While the *GSR* dataset catalogs civil unrest developments, this work also experimented with *GDELT* [15], which is a more comprehensive database of events. It covers most regions of the world in more granular categories, many of which have a violent nature. One example of a *GDELT* event is an occurrence of *ethnic cleansing* on January 24, 2005, by Iraqi forces on individuals of Iranian origin. The event took place in latitude 31.0914 and longitude 46.0872, in the *Dhi Qar* province of *Iraq*. In this study, facts of this nature are used in the generation of storylines, in the calculation of their *SAI* values to nearby regions, and in the verification of whether *GDELT* matched other similar events for the regions of highest *SAI* values. If it did, the events were correctly associated for that particular *SAI* value. Here, a subset of *GDELT* events (the *Observed Events*), is used in the calculation of the *SAI* for a region, and then a different set of *GDELT* events (the *Unobserved Events*), are used in the calculation of recall as explained further below.

Table 5 Recall results based on 5.1 million *GDELT events* in four different categories

GDELT event type	Source country[a]	Target Country (Storyline)	Observed Events / Unobserved Events	Avg. SAI	Recall
THREATEN (political dissent, repression, military force, occupation, attack, mass violence)	AFG	Iran (s_1)	8,245 / 14,465	3.15	0.57
		Pakistan (s_2)	7,129 / 15,842	1.75	0.45
PROTEST (political dissent, rally, hunger strike, passage obstruction)	IRN	Afghanistan (s_3)	5,745 / 9,575	2.03	0.60
		Iraq (s_4)	6,054 / 9,924	2.43	0.61
		Pakistan (s_5)	5,347 / 7,638	1.90	0.70
		Turkey (s_6)	2,118 / 5,573	2.71	0.38
COERCE (seize property, impose sanctions, ban political parties, enact martial law, arrest)	IRQ	Iran (s_7)	10,218 / 12,615	3.21	0.81
		Kuwait (s_8)	3,151 / 5,626	0.60	0.56
		Syria (s_9)	7,211 / 11,093	2.74	0.65
		Turkey (s_{10})	1,616 / 3,298	1.12	0.49
ASSAULT (hijacks, torture, killings, suicide bombings)	PAK	Afghanistan (s_{11})	2,744 / 3,563	2.98	0.77
		India (s_{12})	5,091 / 20,364	3.43	0.25
		Iran (s_{13})	144 / 182	2.45	0.79

Each event type was investigated from a source country to a target country based on initial storylines s_1 through s_{13}. The observed events were included in the calculation of the *Avg. SAI* scores. Recall is the percentage of the unobserved events that had an *SAI* score equal to or greater than the average *SAI* score

[a] Afghanistan, Iran, Iraq, Pakistan

s_1: TALIBAN capture MAZHAR-I-SHARIF occupy IRANIAN CONSULATE kill DIPLOMATS.

s_2: STUDENTS protest FORCES kill QASIM KHAN secure BORDER.

s_3: TEHRAN hosts REFUGEES clash POLICE threaten ECONOMY.

s_4: AIRCRAFT fire MISSILE hit STARK kill PERSONNEL.

s_5: AGENTS kills PAKISTANIS chasing GUARDS reported FISHING.

s_6: IRAN starts OIL supply TURKEY monitor BLAST.

s_7: U.S. warns IRAN fight ISRAEL destroy WEAPONS.

s_8: BA fly KUWAIT seize CITY hold PASSENGERS.

s_9: IRAQ accuse SYRIA plan BOMBING rock MINISTRY.

s_{10}: KADEK wins ELECTION combat PKK declares CEASE-FIRE.

s_{11}: NATO attack SALALA engage CHECKPOST wound SOLDIERS.

s_{12}: MUMBAI conspire PAKISTAN deprive EXTREMIST enter HOTEL.

s_{13}: OFFICIAL shot MAN ran BALUCHISTAN taken NARCOTICS.

Table 5 lists four *GDELT Event Types* documented for several countries. The first row, for instance, indicates 8,245 THREATEN-type events perpetrated by an actor[2] in *Afghanistan* (AFG) on an actor in *Iran*. Starting from an initial storyline (shown on the bottom of the table) that took place in the *Source Country*, the *SAI* values to each of the *Observed Events* in the *Target Country* are calculated. Thus for row 1, we use S_i to calculate the *SAI* values to all the 8,245 observed events in *Iran*, which yielded an average *SAI* of 3.15. Recall is

[2] An actor can be a political organization, the military, militias, terrorist organizations, and individuals, among others.

then the percentage of the *Unobserved Events* that had an *SAI* value of x=3.15 or more, which can be generalized as **recall-2** $= \frac{|SAI(unobservedevents) \geq x|}{|unobservedevents|}$. The *SAI* values can also be compared, allowing one to state that violent events between AFG and *Iran* in row 1 was more probable than violent events between AFG and Pakistan in row 2, where the *Avg. SAI* was lower(1.75), for that category of events.

At first glance, Table 5 shows that the three rows of highest recall (S_7, S_{11}, and S_{13}) also had relatively high *SAI* values. This type of consistency is highly desirable as it may signal that violent events in areas of high *SAI* values have a high potential to be identified and thus associated correctly into the analysis. This consistency, however, must be interpreted carefully as high *SAI* values do not necessarily imply high recall. This is the case with S_6, which were PROTEST-related events between *Iran* and *Turkey*. Its recall value was poor (0.38) because most of the 5,573 unobserved events were not in the PROTEST category. It would be unwise to assert a successful association of events for those. A similar scenario can be seen for S_{12}, where the events between *Pakistan* and *India* were of various natures. One lesson to be learned here is that distribution of event types is an important factor. It is important to filter out storylines that are completely different from the domain in question.

In terms of association analysis, *SAI* operates on Big Data as a criteria to differentiate highly likely events from improbable ones. Here, different observations can be made. The first storyline (S_1) tells about a *TALIBAN* attack on a *CONSULATE* affecting *DIPLOMATS*. Since 57 % of unobserved events that have similar entities are recalled, it can be asserted a 57 % chance that an event with those entities will reoccur in a nearby location. Looking down Table 5, one can make other associations, such as a 65 % chance of an *Iraq*-led attack on a *Syria* target, as examplified in S_9. Indeed, several of such events can be indentified, such as a *Taliban* attack on the *U.S. Consulate* in 2010, and a militia-led suicide bombing by an Iraqi national in Syria in 2011. Table 5 also indicates that the regions between *Pakistan* and *India* provided the best chances for succesfull association of events in the category of *ASSAULTS*, since these two regions had the highest *SAI* values for that category (3.43). The same is true for *Iran* and *Turkey* for the category of PROTEST (2.71) and *Iraq* and *Iran* for COERCE (3.21). Figure 8 depicts spatial propagation of events based on the

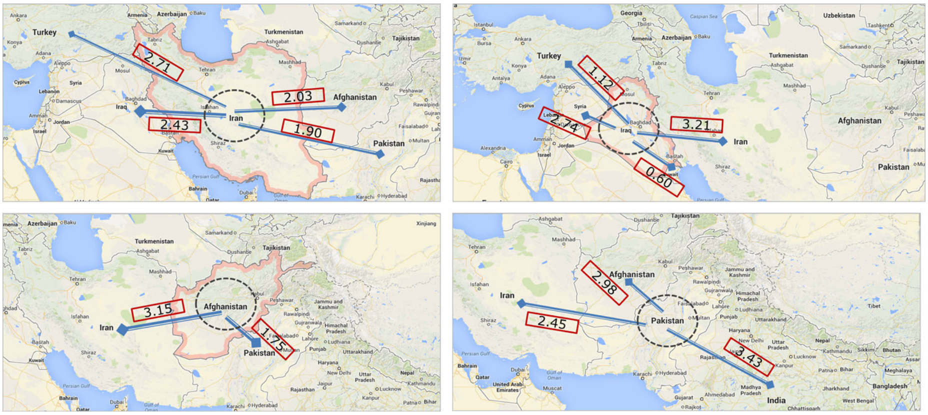


Fig. 8 Spatial propagation of violent events in four parts of Asian countries enclosed in circles. Clockwise from the top left, event types include protests originating in *Iran*, coercion in *Iraq*, assaults in *Pakistan*, and threats in *Afghanistan*. The *boxed* numbers represent the *SAI* scores between the source country and the country pointed to. Higher *SAI* values indicate higher potential for a successful forecast

four *Source Countries* of Table 5 using their corresponding storylines. Visually, higher *SAI* values indicate better chances of a successful association. Values can always be compared starting from the same source country propagating to others, or constrasted across different sources and different destinations.

5.3 Association with Spatio-logical inference

This section applies *spatio-logical inference* to transform storylines into weight-based rules, which are then used in the association analysis.

The dataset comprised 3.2 million *GDELT* violent events that took place in *Afghanistan*. Out of those records, 2.1 million were used to extract rules, find events of high probability of occurrence using *Spatio-Logical Inference*, and use the results to find the number of similar events that exist in the remaining 1.1 million. The measures were: **recall-3**$= \frac{similaridentified}{similaridentified+similarmissed}$ as the number of similar events that were identified over the total number of similar events among the 1.1 million; **precision-1**$= \frac{similaridentified}{allretrieved}$ as the number of similar events that were identified over all retrieved records. Similar events denote events of the same ontological resolution (specified by the user) located within *distance to satisfaction* $\leq t$ of one another, where t is a threshold. The experiments evaluated different distance thresholds.

To extract rules from the dataset, Algorithm 1 was used, for which a brief example is given here. Consider the three *GDELT* event types shown in Table 6 and geolocated in the corresponding image, which has *Afghanistan* as the region of study. The frequency for each event type is shown in parenthesis. Because the two closest events are 🅐 and 🅑, at a distance of 115 km, these two events make up the body of the rule. The remaining one, event 🅒, becomes the implication:

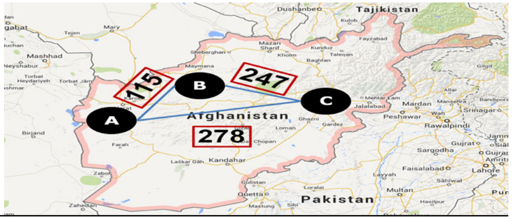
carryout-vehicular-bombing(AFGMOS,AFGREB) ∧ **use-as-human-shield** (AFGREB,AFGCVL) $\Longrightarrow$ **attempt-to-assassinate**(AFGCVL,AFGMIL)

To add the *soft truths*, look at Table 6 and see that the probability of event 🅐 $= \frac{15}{45} = 0.33$, 🅑 $= \frac{5}{45} = 0.11$, and 🅒 $= \frac{25}{45} = 0.55$. The overall weight of the rule is the average distance between the three events, normalized in the range [0,1], which can be calculated as 0.76, assuming a minimum distance of 0 km, and a maximum distance of 278 km. Thus the final rule looks like:

$$0.76 : \overbrace{carryoutvehicularbombing(AFGMOS, AFGREB)}^{0.33} \wedge \overbrace{useashumanshield(AFGREB, AFGCVL)}^{0.11}$$
$$\Longrightarrow \overbrace{attempttoassassinate(AFGCVL, AFGMIL)}^{0.55}$$

The above rule is then used to find its *distance to satisfaction* as described in Section 4.2. In the experiments, the overall weight of every rule is set to 1.0 (every rule is equally important), and thus, the focus is set on the soft truths instead. The correctly associated events are the rules with the least distance to satisfaction. Based on that, *precision* and *recall* are utilized as the evaluation measures. It should be clear that the above example is a simple scenario with only three events. Given vast numbers of events, the number of rules can easily explode. Optimizations should be done, such as shortening distances or filtering out specific event types in order to alleviate computation costs. In this study, weights are based on

Table 6 Example of three *GDELT* events located in different areas of *Afghanistan* in 2011

	Event description (instances)	Source[a]	Target[a]	Lat	Lng	
A	Carry out vehicular bombing (15)	AFGMOS	AFGREB	34.3333	70.4167	 A B C 115 247 278 Afghanistan Tajikistan Pakistan Dushanbe Mashhad Kandahar Peshawar Rawalpindi Quetta Multan Lahore
B	Use as human shield (5)	AFGREB	AFGCVL	34.5167	69.1833	
C	Attempt to assassinate (25)	AFGCVL	AFGMIL	32.3472	68.5932	

The number in parenthesis is the total number of events of that kind reported in *Afghanistan* for that year. The latitude and longitude values represent the specific points where one instance of that event took place between the source and target groups. The image shows the distance in km between the different events, which are used to generate rules for *Spatio-logical Inference*

[a] AFG=Afghanistan, MOS=Muslim group, REB=Rebel group, CVL=Civilians, MIL=Military

frequencies and spatial distances, but it is possible that different approaches may be better suited for different domains of knowledge.

Discussion In the context of violent events, a key consideration is whether relevant events can be associated with one another, knowing that relevance is a highly subjective matter. For measurement purposes, this paper defines relevance in a comparative scale based on either *Euclidean* distance or *distance to satisfaction*: lower values are always more relevant than higher ones. Association among events can be investigated in three configurations: (1) all events are the same, such as when instances of fights result in other fights; (2) all events are different, such as when a fight and police crackdown result in a riot; (3) otherwise, events are mixed. Assume that there exists a set of *trigger events* (ϕ_1 to ϕ_n) that lead to a *final violent event* (ϕ_{final}) with a d *Euclidean* distance or *distance to satisfaction*. Then one can assert a successful association for other unseen *final violent events* provided that the *trigger events* lead to the same *final violent events* with the same or lower distance d. In other words, comparing the association between two sets of events, if the events match (or partially match) on at least one *trigger event* and distance is just as low, then a successful association is made. If no events match or distance d is off (higher than a threshold), then the association is a miss.

Using the above ideas, all events from our dataset that fit those conditions are retrieved, allowing a count of how many were associated correctly, and how many were not. For simplicity, the region of study is limited to a range where the maximum distance between any two events is 100 km. Figure 9 shows six plots with different measurements for discussion. High recall values indicate that previously-unseen events were being found without going over a distance limit. This is shown in Fig. 9a, in which recall values ranged from 40 % to 66 %. In the range where distance between events lay between 0 to 50 km, recall remained fairly constant at around 62 % for mixed events. This indicated that, for many of the generated rules, their constituent events lead to the same *final violent event* with a distance of 50 km or less. For events of the same type, recall trended upward up to 50 km, but only got worse thereafter. More intriguing were unique events, in which recall was good with short distances (0–20 km) or long distances (80–100 km), but often worse for distances between (21–79 km). The lesson learned from this example was the following: the *soft truth* values established in the rules seemed to be appropriate for the initial part of the graph (shorter distances) and the late stages (longer distances), but may not have been ideal for mid-range distances.

Those values were candidates for adjustment, but trying to readjust them may only be a temporary fix, since Big Data cannot be assumed to have a specific distribution. Proceeding to Fig. 9c, the *distance to satisfaction* trended down most of the way with the exception of a spike at 0.6. The downward portion related to the notion that fewer of the *final violent events* were being found, or when found, the *distance to satisfaction* was too high (i.e., above the limit established by the rule that found it). The analyst would be interested in investigating the events associated with low recall to see if adjusting the *soft-truth* values would afford better results. It is possible that the values were indeed correct, and that the low recall came as a result of violent events in the unseen data not matching the ones in the observed data.

Similar trends as the above were also seen in Fig. 9b, which shows precision by *Euclidean* distance. In general, one would expect high recall for short distances and vice-versa. Intuitively, government in Kabul experienced many bombings over time, but ones which were not necessarily related to other bombings in far-away cities, such as Charikar. However, the data indicated that, in many instances, longer distances between events displayed higher precision than shorter ones. This is the case in Fig. 9b where the highest

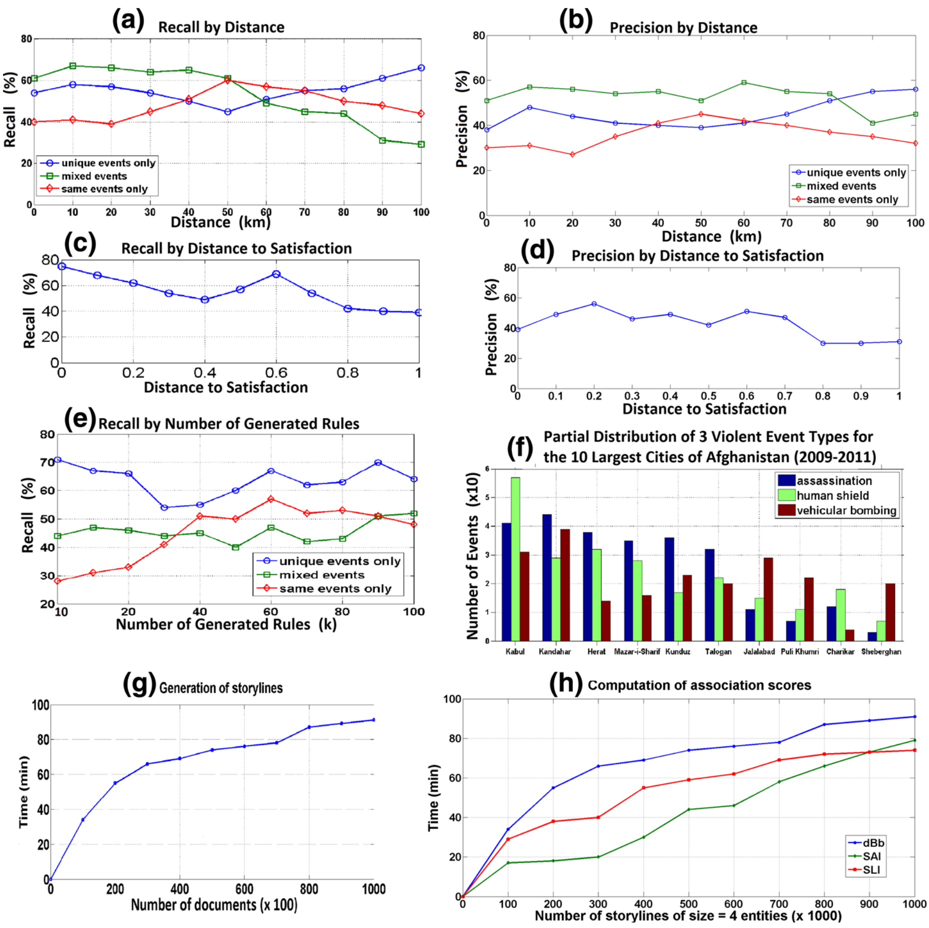


Fig. 9 Results from *spatio-logical inference*. **A** Effect of distance between events on recall. **B** Effect of distance between events on precision. **C** Effect of the *distance to satisfaction* on recall. **D** Effect of the *distance to satisfaction* on precision. **E** Effect of the number of generated rules on recall. **F** Distribution of violent events in cities of Afghanistan. **G** Effect of the number of documents on storyline generation. **H** Time it takes to compute association scores based on the number of storylines

precision for mixed events was approximately 60 % with a distance of 60 km. For unique events, this fact was even more pronounced, since the highest precision (57 %) lined up with the longest distance (100 km). This is indicative of a particular type of event that takes place in many locations (e.g., protests against corruption taking place across multiple cities): the violent events matched with similar conditions (i.e., similar *trigger events*) even when the cities were far apart. Figure 9d shows the effects of *distance to satisfaction* on precision. This trend did not deviate significantly from the *Euclidean* distance approach, even though high precision at times does come from lower distances. The fact that the two approaches had similar results is encouraging because it indicates that our reasoning was valid.

In terms of processing times, Fig. 9g provides a snapshot of how long it takes to process documents into storylines. Documents, in this case, are tweets where the storyline generation process was distributed in a *Hadoop* cluster of 10 machines (4 GB Ram, quad-core, 64-bit). At times, the generation of storylines is close to linear with respect to the number of documents. Most of the cost of building storylines comes from identifying entities (people, organizations, objects), geocoding, and indexing them. Where the graph grows relatively fast, such as between 0 and 20,000 documents, indicates that not many repeated entities and locations are being found. This requires more processing, which not surprisingly takes longer.

Flat trends, such as between 40,000 and 60,000 documents, point to repeated data (such as locations that have already been identified previously) that do not need to be processed again. With repeated data, more documents can be handled. In a Big Data scenario of a specific domain, this graph may trend flatter, since specific domains will rely on a lesser variation of data than general data would. Figure 9h provides a snapshot of how long it takes to go through the computation of the association scores, which are the *DbB*, *SAI*, and SLI. This is based on 1,000,000 storylines composed of at most 4 entities. The general trend is that *SLI* can complete more calculations than *SAI* or *DBb* for the same amount of time. This is due to the fact that *SLI* is mostly about table lookups for historical frequencies along with simple calculations. On the other hand, *SAI* needs to look for storylines in neighboring regions and identify similar entities. Thus, it tends to take longer to compute its association score. Note, however, that these graphs do not relate to the "goodness" of the storylines. It only provides an idea of how long it takes to run each method based on an increasing number of documents. For informational purposes, 1 million tweets can be translated into 4-entity storylines (and calculation scores computed) in approximately 90 min, in a distributed environment of 10 machines (Hadoop). In a standalone machine, the same process takes approximately 7.5 hours. For Big Data, this is an important consideration, but which must be analyzed carefully in terms of storyline coherence.

The above discussion points to the importance of relating event types, locations, distances, and frequencies in the discussion of violent events. These were the components used in the generation of the event-based rules. These features represent a robust set of components that populate Big Data. Figure 9e summarizes recall in terms of the number of generated rules according to event type. This time, distance was disregarded, which had a different effect on the results. When distance was not considered, recall was consistently high when events had different types, but suffered considerably for mixed ones, with a higher variation for same event types. The closest that the three lines came together was at approximately 33k generated rules, where recall ranged from 41 % to 54 %. This is a significant difference from the distance approach, and underscores the importance of spatial analysis. For illustrative purposes, Fig. 9f depicts the distribution of three events for the 10 largest cities in *Afghanistan*, which were used in this dataset. It shows, for instance, that (for this partial dataset) *vehicular bombings* were mostly frequent in *Kandahar*, *Kabul*, and *Jalalabad* (in this order), while *Kabul* itself saw most of the *human shield* events. While this graph is not the complete dataset used in the experiments, it gives the reader a sense of the spatial locations being investigated and the event types being investigated.

Finally, Table 7 displays some of the events that the proposed approach was able to correctly associate. Starting from a sample generated rule ($G1$), whose *final violent event* was related to *destruction of property*, and had a *distance to satisfaction* $= 0.25$, the table first shows a set of four rules that were correctly associated ($F1$, $F2$, $F3$, $F4$). $F1$, for example, tells about some sort of "negotiation" that involves an action of "release", which eventually ended up as "confiscation of property". Without the benefit of external knowledge, the details of this case was not known. However, it can be affirmed with confidence that this event was very close in concept to the original rule $G1$, which also had a "release" component, involved "destruction of property", and had lower *distance to*

Table 7 Examples of events that were correctly associated with one another or missed based on the generated rule shown across the top row

	Generated rule	Distance to satisfaction: 0.25
$G1$	engage(AFGGOV,RADMOS) $\wedge$ demand-release(AFG,COP) $\Longrightarrow$ destroy-property(AFGREB,RADMOS)	
	Events correctly forecast by rule $G1$	
$F1$	halt-negotiation(AFGCOP,UAF) $\wedge$ demand-release(AFG,COP) $\Longrightarrow$ confiscate-property(AFGGOV,RADMOS) (0.13)	
$F2$	engage(AFGGOV,RADMOS) $\wedge$ impose-embargo(AFGSPY,AFGCRM) $\Longrightarrow$ seize(AFGGOV,AFGINSTALUAF) (0.17)	
$F3$	cooperate-militarily(AFGCOP,AFG) $\wedge$ impose-curfew(AFGSPY,AFGCVL) $\Longrightarrow$ destroy-property(AFGGOV,RADMOS) (0.12)	
$F4$	ban parties(AFGCOP,UAF) $\wedge$ demand-material-coop(AFGGOVBUS,AFGCVL) $\Longrightarrow$ destroy-property(AFGREL,RADMOS) (0.24)	
	Missed forecasts	Reason for miss
$M1$	engage(AFGGOV,RADMOS) $\wedge$ reject(AFG,AFG) $\Longrightarrow$ mobilize-armed-forces(AFG,RADMOS) (0.20)	wrong *final violent event*
$M2$	halt-negotiation(AFGCOP,UAF) $\wedge$ use-tactics-violent(AFG,COP) $\Longrightarrow$ destroy-property(AFGGOV,RADMOS) (0.20)	no match on *trigger events*
$M3$	expel(AFGMIL,AFGELI) $\wedge$ rally-opposition(AFGGOV,AFGREF) $\Longrightarrow$ demand-release(AFGGOV,RADMOS) (0.38))	high *distance to satisfaction*
$M4$	engage(AFGEDU,AFGMIL) $\wedge$ reduce-econ-aid(UAF,AFGREF) $\Longrightarrow$ destroy-property(AFGGOV,RADMOS) (0.31)	high *distance to satisfaction*

The final violent event is **destruction of property** as shown in the implication of the generated rule $G1$. For each missed association, a reason explains why it was not considered valid. The number in parenthesis at the end of each rule denotes *distance to satisfaction*

AFG= Afghanistan, BUS= business, COP=police force, CRM=criminal, CVL=civilian, ELI=elites, GOV=government, MIL=military, MOS=muslim, RAD=radical, REB= rebels, REF=refugee, SPY=spy, UAF=unidentified armed force *satisfaction* than the original rule $G1$ (0.13 as opposed to 0.25). These events took place in 2010 in Afghanistan at a distance of 34 km from each other. The same was true for $F3$, which also dealt with "destruction of property", though coming from totally separate *trigger events* related to "military cooperation" and a "curfew". $F2$ and $F4$ had slightly higher *distance to satisfaction*, albeit still below the limit of 0.25 established by $G1$.

Further down, the table shows four other rules that were not considered valid associations. The first one, $M1$, did not have a similar *final violent event* to $G1$. $M2$ shared no *trigger events* at all with $G1$, and thus was not valid because the approach needs at least one element in common. $M3$ and $M4$ were both too distant in terms of *distance to satisfaction* from 0.25, and thus were rejected as well.

In *storytelling*, the high number of entities and events is always of concern, especially in a Big Data environment. It is important, thus, to understand the number of rules that are generated and how they affect recall, which is shown in Fig. 9e. The plot separates whether the events considered were of the same nature (e.g., bombing followed by another bombing), unique natures (e.g., bombing followed by an assassination attempt), or a mix of them. When event types are mixed, recall remained fairly constant despite the increase in the number of generated rules. It hints at the distribution of the data: events were well spread out throughout space. An analyst studying many event types concurrently may find this fact interesting. The situation was vastly different when the events were all the same or all different. In this case, recall displayed greater variation (28–57 % and 54–71 %, respectively). The graph also shows that fewer rules were not necessarily better than more rules (as one might expect). In fact, some of the best recall values can be seen exactly at the end of the graph when the number of generated rules hits 100 k. Although this may not seem intuitive, violent events can be more easily explainable when their constituent developments have different natures, as opposed to when they are composed of the same event types.

5.4 Comparison of the different association strategies

In this section, the three association strategies explained earlier are put in perspective. The goal is not to find the best strategy, but rather to contrast them. One line of research complementary to this work, but which often does not include spatial *storytelling*, is event detection, to which further reading is suggested. [16, 30]. The following discussion is framed in terms of *precision* and *recall*, as done before.

Table 8 lists a set of 5 event types, labeled $E1$ through $E5$, from the *GDELT* dataset that were targeted as *final violent events*. 2.58 million records were used: 1.8 million as input and 0.78 million for validation. For each event type, the table shows precision and recall values, calculated as explained earlier, using the three technical approaches discussed in Section 3.1. The highest values are shown in bold type. Events considered were those whose probability of occurrence was 10 % or more (less than 10 % was less significant in our dataset).

xThe way to interpret the table, exemplified for row 1, is as follows. Upon running *Distance-based Bayesian Inference* for event $E1$ (*attempt to assassinate*) in the initial set of 1.58 million events, the results indicated 5,101 combinations (not shown in table) of *trigger events* that led to $E1$ with a probability $\geq$ 10 %. However, when validating against the remaining 0.78 million records, those combinations only contained 985 out of 2662 events with a probability$\geq$ 10% (and that shared at least one event with the generating combination), yielding a precision of 0.37. For recall, 985 combinations were found, but 1539 should have been identified, resulting in a recall of 0.64. For the other approaches,

Table 8 Comparison of precision and recall for three different approaches: *Distance-based Bayesian Inference*, *Spatial Association Index*, and *Spatio-logical Inference*

	Distance-based Bayes (DbB)			Spatio-logical Inference (SLI)			Spatial Association Index (SAI)		
Event	precision	recall	f-measure	precision	recall	f-measure	precision	recall	f-measure
E1-attempt to assassinate	0.35	0.81	0.48	0.51	0.58	0.54	0.55	0.77	0.64
E2-carry out vehicular bombing	0.61	0.75	0.67	0.57	0.72	0.63	0.66	0.80	0.72
E3-engage in violence	0.48	0.71	0.57	0.54	0.72	0.61	0.66	0.79	0.71
E4-conduct strike or boycott for rights	0.45	0.65	0.53	0.78	0.61	0.68	0.67	0.79	0.72
E5-destroy property	0.51	0.70	0.59	0.61	0.77	0.68	0.45	0.52	0.48

For each row, the highest values are shown in red letters instead of a simple probability, the criteria were $SAI \leq 1.0$ and *distance to satisfaction* $df \leq 1.5$.

Discussion: For *Distance-based Bayes*, precision was often low, but with mixed signals. It was significant for $E2$ (carry out vehicular bombing), but much lower for $E1$ (attempt to assassinate) and for $E4$ (conduct strike or boycott for rights). The reason for the low scores had to do with frequency, which was fairly low for this type of event. In general, recall was consistently high, especially in $E1$, whose frequency of "assassinations" was also high in the dataset.

The *Spatial Association Index* demonstrated the highest precision of any of the approaches for events $E1$, $E2$, and $E3$. Its positive aspect was consistency even when it was not the highest. Its precision was never lower than 0.45, though for larger distances, lower scores were observed consistently. SAI is highly sensitive to how many events are far apart versus nearby, and seems to favor the latter. The data distribution was certainly a factor here as was event colocation. For example, the dataset had many instances of the same pairs of events that led to $E4$ (*conduct strike or boycott for rights*) with high values of SAI. This explains the 0.67 precision of $E4$. It is true that some of the SAI values were low ($E5$), but those can be explained by the low numbers of similar events in the validating dataset. Its recall values were good across all events with the exception of $E5$ (0.52).

Two observations can be made about *Spatio-logical Inference* (SLI): first, precision showed good consistency for low *distances to satisfaction*, which is desirable in terms of association. However, one should also expect low precision for high *distances to satisfaction*, which in general does not occur when df >1.0. While high precision is normally desirable, it would be preferable for df to oppose precision hand-in-hand (low to high, and high to low). Precision values were unexpectedly high for $E4$ and $E5$ (0.78 and 0.61, respectively). Indeed, these values came from many rules that were established by far-apart events of the same ontological category with high *soft truth* values, and thus their high precision. *SLI* behaved in a stable manner in terms of recall, and interestingly, especially when distances were long. While Table 8 only shows a limited number of results, the overall experience points to SLI as having the best recall results.

5.5 Key Observations

The discussion in this paper evolved from storylines composed of events, and possessing two of the challenges that come with Big Data: high data volumes and high data variability. As such, the first inclination is to favor the three distance-based approaches (DbB, SAI, and SLI), and momentarily disregard methods that do not take into account location as a feature. It would be attractive to single one of them out as the most robust associating strategy, one which would be able to capture all associations with high certainty. While such an answer is not feasible, several considerations can be made based on knowledge of the dataset and the adjustment of parameters:

1. **Distance variation**: The experiments of Section 5.4 demonstrated that, for datasets across large spatial regions (across countries, for example), *DbB* provided higher precision than the other approaches. For low event distances (for example, crime hotspots in Washington D.C.), on the other hand, either SAI (Section 5.2) or SLI (Section 5.3) showed better performance.
2. **Precision vs. recall**: When high recall is more important than high precision, better results can be obtained with DbB or SFI. On the other hand, SAI showed higher precision scores in experiments not shown here, as long as the application was constrained

to a specific domain. For general use, however, there is still little basis to advocate for one method versus the others.

3. **Concept resolution**: Categorization can contribute significantly to precision and recall. Combining several events into common classifications increases the chances of finding similar occurrences, and helps deal with the high variability of Big Data. An important implication, however, is that such combinations result in loss of information, and must be taken carefully.
4. **Data distribution**: The characteristics of the data, such as distribution, should be investigated. Since a single distribution type cannot be guaranteed under Big Data, one way to avoid this problem is to apply sampling during the storyline generation process. Many sampling methods have been proposed, which can potentially make the storylines less prone to bias and and lead to better application results.
5. **Thresholds**: The experiments of Section 5.4 compared the different methods based on midpoint thresholds, such as a minimum probability of 10% that an event would occur. These are values that worked well in the scope of this paper, but can certainly be manipulated or even parameterized as user-defined inputs. These distances are highly dependent on the application domain and, whenever possible, should be experimented with until optimal values are identified.

6 Conclusion

This study demonstrates that spatio-temporal storytelling is able to capture important associations among violent events reported in social media and traditional databases, two common sources of Big Data. The major contributions are three methods of association analysis: *Distance-based Bayesian Inference* relates similar events that are described differently, addressing high variability in Big Data; *Spatial Association Index* measures the influence of the storylines from one geographical location to others, limiting high data volumes to constrained regions; and *Spatio-logical Inference* compute a score to determine if a set of initial events is related to a final violent event, filtering irrelevant developments. The latter two provide a means to deal with the high volumes in Big Data. These methods can be highly valuable in the analysis of event searches, propagation, influence, and forecasting.

References

1. Bolzoni P, Helmer S, Wellenzohn K, Gamper J, Andritsos P (2014) Efficient itinerary planning with category constraints. In: Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL ’14. ACM, NY, USA, pp 203–212. doi:10.1145/2666310.2666411
2. Bouros P, Sacharidis D, Bikakis N (2014) Regionally influential users in location-aware social networks. In: Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL ’14. ACM, NY, USA, pp 501–504. doi:10.1145/2666310.2666489
3. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30:107–117
4. Chan J, Bailey J, Leckie C (2008) Discovering correlated spatio-temporal changes in evolving graphs. Knowl Inf Syst 16:53–96
5. Chan J, Bailey J, Leckie C (2009) Using graph partitioning to discover regions of correlated spatio-temporal change in evolving graphs. Intell Data Anal (IDA) 13:755–793
6. George B, Kang J, Shekhar S (2009) Spatio-temporal sensor graphs (stsg): a data model for the discovery of spatio-temporal patterns. Intell Data Anal (IDA) 13:457–475

7. Hossain MS, Andrews C, Ramakrishnan N, North C (2011) Helping intelligence analysts make connections. In: Workshop on scalable integration of analytics and visualization, AAAI '11, pp 22–31
8. Hossain MS, Butler P, Ramakrishnan N, Boedihardjo A Stortytelling in entity networks to support intelligence analysts. In: Conference on Knowledge Discovery and Data Mining (KDD'12), pp 1375–1383
9. Hossain M. S., Gresock J., Edmonds Y., Helm R., Potts M., Ramakrishnan N. (2012) Connecting the dots between pubmed abstracts, vol 7
10. Iarpa - open source indicators program (osi) (2014). http://www.iarpa.gov/solicitations_osi.html
11. Kimmig A, Bach SH, Broecheler M, Huang B, Getoor L (2012) A short introduction to probabilistic soft logic. In: NIPS Workshop on probabilistic programming: Foundations and applications
12. Kleinberg J. (1998) Authoritative sources in a hyperlinked environment. In: Society of industrial and applied mathematics (SIAM), pp 668–677
13. Kreinovich V, Kosheleva O (2008) Computational complexity of determining which statements about causality hold in different space-time models. Theor Comput Sci 405(1-2):50–63
14. Kumar D, Ramakrishnan N, Helm RF, Potts M (2008) Algorithms for storytelling. IEEE TKDE 20(6):32. http://doi.ieeecomputersociety.org/10.1109/TKDE.2008.32
15. Leetaru K., Schrodt P. (2013) Gdelt: Global database of events, language, and tone, 1979-2012. In: Proceedings International Studies Associations Annual Conference (ISA)
16. Li Z, Wang B, Li M, Ma WY (2005) A probabilistic model for retrospective news event detection. In: ACM SIGIR Conference on research and development in information retrieval, SIGIR '05, pp 106–113
17. Liu M, Fu K, Lu CT, Chen G, Wang H (2014) A search and summary application for traffic events detection based on twitter data. In: Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '14. ACM, NY, USA, pp 549–552. doi:10.1145/2666310.2666366
18. Marchiori M (1997) The quest for correct information on web: Hyper search engines. In: World wide web conference (WWW), pp 1225–1235
19. Mondo G. D., RodrGuez M., Claramunt C, Bravo L, Thibaud R (2013) Modeling consistency of spatio-temporal graphs. Data Knowl Eng 84:59–80
20. P. Mohan S, Shekhar JS, Rogers J (2012) Cascading spatio-temporal pattern discovery. Trans Knowl Data Eng (TKDE) 24(11):1977–1992
21. Radinsky K, Davidovich S, Markovitch S (2012) Learning causality for news events prediction. In: World wide web conference (WWW), pp 909–918
22. Radinsky K, Davidovich S, Markovitch S (2012) Learning to predict from textual data. J Artif Intell Res (JAIR) 45:641–684
23. Radinsky K, Horvitz E (2013) Mining the web to predict future events. In: Conference on web search and data mining, WSDM '13, pp 255–264
24. Santos RD, Shah S, Chen F, Boedihardjo A, Butler P, Lu CT, Ramakrhishnan N (2016) A framework for intelligence analysis using spatio-temporal storytelling. Geoinformatica, Int J Adv Comput Sci Geogr Inf Syst:1
25. Shahaf D, Guestrin C (2010) Connecting the dots between news articles. In: ACM Conference on knowledge, discovery, and data mining (KDD '10), pp 745–770
26. Shahaf D, Guestrin C, Horvitz E Metro maps of science. In: Conference on Knowledge Discovery and Data Mining, KDD'12, pp 1122–1130
27. Shahaf D, Guestrin C, Horvitz E Trains of thought: Generating information maps. In: World Wide Web Conference, WWW'12, pp 899–908
28. Shekhar S, Chawla S (2003) Spatial databases: a tour. Prentice Hall, New York
29. Turner S (1994) The creative process: A computer model of storytelling and creativity. Psychology Press, pp 122–123
30. Vavliakis KN, Symeonidis AL, Mitkas PA (2013) Event identification in web social media through named entity recognition and topic modeling. Data Knowl Eng 88:1–24
31. Wang B, Wang X (2011) Spatial entropy-based clustering for mining data with spatial correlation. In: Proceedings of the 15th pacific-asia conf. on adv. in knowledge discovery and data mining, PAKDD'11, pp 196–208
32. Zhang J. D, Chow C. Y, Li Y (2014) Lore: Exploiting sequential influence for location recommendations. In: Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '14. ACM, NY, USA, pp 103–112. doi:10.1145/2666310.2666400
33. Zhou X, Chen L (2014) Event detection over twitter social media streams. VLDB J 23(3):381–400. doi:10.1007/s00778-013-0320-3

Raimundo Dos Santos received a Bachelor's Degree in Computer Science from the University of South Florida in 1999. He received a Master's and PhD degrees in Computer Science from Virginia Tech in 2010 and 2014, respectively. He has published in several venues including ACM-GIS, IEEE-ICTAI,IJTAI, and Geoinformatica. His research focuses on semantic entity analysis and Spatial Data Management, including retrieval, exchange, and processing of information for Geographic Information Systems and location-based services. Other interests include data integration, graph mining, and analytical methods for semantic storytelling.

Arnold P. Boedihardjo received his BS degrees in Mathematics and Computer Science from Virginia Tech in 2001. He received his MS and PhD degrees in Computer Science from Virginia Tech in 2006 and 2010, respectively. He has published in various scholarly venues such as IEEE International Conference on Data Engineering, IEEE International Conference on Data Mining, ACM Conference on Information and Knowledge Management, Knowledge and Information Systems Journal, and IET Communications Journal. His research interests include data stream systems, spatial databases, information retrieval, optimizations, networking, and statistical learning. He is currently a research scientist at the U.S. Army Sumit Shah received a BS and MS in Computer Science from Virginia Tech. He has worked extensively in software engineering and systems architecture and is currently a PhD. candidate in the Department of Computer Science at Virginia Tech. He has published work at ACM GIS and other academic venues. His research focuses on large scale data mining, Big Data, information retrieval, and location-based services. Other interests include mobility and data visualization.

Sumit Shah received a BS and MS in Computer Science from Virginia Tech. He has worked extensively in software engineering and systems architecture and is currently a PhD. candidate in the Department of Computer Science at Virginia Tech. He has published work at ACM GIS and other academic venues. His research focuses on large scale data mining, Big Data, information retrieval, and location-based services. Other interests include mobility and data visualization.

Feng Chen is an assistant professor at State University of New York at Albany. He received his B.S. from Hunan University, China, in 2001, M.S. degree from Beihang University, China, in 2004, and Ph.D. degree from Virginia Polytechnic Institute and State University in 2012, all in Computer Science. He has published 25 refereed articles in major data mining venues, including ACM-SIGKDD, ACM-CIKM, ACMGIS, IEEE-ICDM, and IEEE-INFOCOM. He holds two U.S. patents on human activity analysis filed by IBM's T.J. Watson Research Center. His research interests are in the areas of statistical machine learning and data mining, with an emphasis on spatiotemporal analysis, social media analysis, and energy disaggregation.

Chang-Tien Lu received the MS degree in computer science from the Georgia Institute of Technology in 1996 and the PhD degree in computer science from the University of Minnesota in 2001. He is an associate professor in the Department of Computer Science, Virginia Polytechnic Institute and State University and is founding director of the Spatial Lab. He served as Program Co-Chair of the 18th IEEE International Conference on Tools with Artificial Intelligence in 2006, and General Co-Chair of the 20th IEEE International Conference on Tools with Artificial Intelligence in 2008 and 17th ACM International Conference on Advances in Geographic Information Systems in 2009. He is also serving as Vice Chair of the ACM Special Interest Group on Spatial Information (ACM SIGSPATIAL). His research interests include spatial databases, data mining, geographic information systems, and intelligent transportation systems.

Naren Ramakrishnan is the Thomas L. Phillips Professor of Engineering at Virginia Tech. He directs the Discovery Analytics Center (DAC; http://dac.cs.vt.edu) at Virginia Tech, a university-wide effort that brings together researchers from computer science, statistics, mathematics, and electrical and computer engineering to tackle knowledge discovery problems in important areas of national interest, including intelligence analysis, sustainability, neuroscience, and systems biology. Ramakrishnan's research has been supported by NSF, DHS, NIH, NEH, DARPA, IARPA, ONR, General Motors, HP Labs, NEC Labs, and Advance Auto Parts. He serves on the editorial boards of IEEE Computer, Data Mining and Knowledge Discovery, and other journals. Ramakrishnan was an invited co-organizer of the NAE Frontiers of Engineering symposium in 2009. He is an ACM Distinguished Scientist (2009).